FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Jan Blechta

Towards efficient numerical computation of flows of non-Newtonian fluids

Mathematical Institute of Charles University

Supervisor of the doctoral thesis: prof. RNDr. Josef Málek, CSc., DSc.
Study programme: Physics
Study branch: Mathematical and Computer Modelling

Prague 2019

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

I declare that this thesis has not been used to obtain another or the same degree.

In .... date .... signature of the author

Věčné památce Ondřeje Blechty

Title: Towards efficient numerical computation of flows of non-Newtonian fluids

Author: Jan Blechta

Institute: Mathematical Institute of Charles University

Supervisor: prof. RNDr. Josef Málek, CSc., DSc., Mathematical Institute of Charles University

Abstract: In the first part of this thesis we are concerned with the constitutive theory for incompressible fluids characterized by a continuous monotone relation between the velocity gradient and the Cauchy stress. We, in particular, investigate a class of activated fluids that behave as the Euler fluid prior activation, and as the Navier-Stokes or power-law fluid once the activation takes place. We develop a large-data existence analysis for both steady and unsteady three-dimensional flows of such fluids subject either to the no-slip boundary condition or to a range of slip-type boundary conditions, including free-slip, Navier's slip, and stick-slip.

In the second part we show that the $W^{-1,q}$ norm is localizable provided that the functional in question vanishes on locally supported functions which constitute a partition of unity. This represents a key tool for establishing local a posteriori efficiency for partial differential equations in divergence form with residuals in $W^{-1,q}$.

In the third part we provide a novel analysis for the pressure convection-diffusion (PCD) preconditioner. We first develop a theory for the preconditioner considered as an operator in infinite-dimensional spaces. We then provide a methodology for constructing discrete PCD operators for a broad class of pressure discretizations. The principal contribution of the work is that a clear and pronounced methodology for dealing with the artifical boundary conditions is given, including the inflow-outflow case, which has not been adequately addressed in the existing literature.

Keywords: non-Newtonian fluids, constitutive theory, fluids with activation, large-data a priori analysis, a posteriori error estimates, preconditioning

Název práce: K efektivním numerickým výpočtům proudění nenewtonských tekutin

Autor: Jan Blechta

Ústav: Matematický ústav Univerzity Karlovy

Školitel: prof. RNDr. Josef Málek, CSc., DSc., Matematický ústav Univerzity Karlovy

Abstrakt: V první části práce se zabýváme konstitutivní teorií nestlačitelných tekutin charakterizovaných spojitým monotónním vztahem mezi gradientem rychlosti a Cauchyho napětím. Speciální pozornost je věnována třídě aktivovaných tekutin, které se před aktivací chovají jako Eulerovy tekutiny, zatímco po aktivaci je jejich odezva stejná jako odezva Navierovy-Stokesovy tekutiny či tekutiny mocninného typu. Pro tuto třídu tekutin je provedena detailní existenční analýza pro velká data k stacionárním a nestacionárním třídimenzionálním prouděním vystavených buď okrajové podmínce nulové rychlosti, či řadě podmínek skluzového typu, včetně volného skluzu, Navierova skluzu a kombinovaného přilnutí-skluzu.

Druhá část se zabývá lokalizací $W^{-1,q}$ normy za předpokladu, že uvažovaný funkcionál se nuluje na fukcích s lokálním nosičem, které tvoří rozklad jednotky. To zvláště dovoluje zajistit lokální aposteriorní efektivitu u parcialních diferencialních rovnic v divergentním tvaru s residuály ve $W^{-1,q}$.

V třetí části předkládáme novou analýzu tzv. PCD (pressure convection-diffusion) předpodmínění. Nejdříve budujeme novou teorii PCD předpodmínění jakožto operátoru v nekonečně-dimenzionálních prostorech. Potom poskytujeme metodiku ke konstrukci diskrétních PCD operátorů pro širokou třídu diskretizací tlaku. Hlavní přínos práce je jasná a zřetelná metodika nakládání s umělými okrajovými podmínkami, včetně situací s nátokem a výtokem, což nebylo doposud v literatuře dostatečně ošetřeno.

Klíčová slova: nenewtonské tekutiny, konstitutivní teorie, tekutiny s aktivací, apriorní analýza pro velká data, aposteriorní odhady chyby, předpodmínění

Contents

Preface **xi**

I On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion[1] **1**

1 Introduction . . . 1

2 Classification of incompressible fluids . . . 3

2.1 Classical power-law fluids . . . 4

2.2 Generalized power-law fluids and stress power-law fluids . . . 7

2.3 Fluids that can be viewed as a mixture of power-law fluids . . . 8

2.4 Fluids with bounded shear rate or bounded shear stress . . . 8

2.5 Activated fluids . . . 10

2.6 Classification of incompressible fluids . . . 13

2.7 Activated boundary conditions . . . 13

3 Mathematical analysis of flows of activated Euler fluids . . . 16

3.1 Function spaces . . . 20

3.1.1 No-slip case . . . 20

3.1.2 Slip case . . . 21

3.2 Analysis of steady flows . . . 22

3.2.1 No-slip case . . . 22

3.2.2 Slip case . . . 26

3.3 Analysis of unsteady flows . . . 28

3.3.1 No-slip case . . . 28

3.3.2 Slip case . . . 35

4 Concluding remarks . . . 37

Appendix A Auxiliary convergence tools . . . 38

Appendix B Examples of maximal monotone graphs . . . 42

References . . . 44

II Localization of the $W^{-1,q}$ norm for local a posteriori efficiency[2] **49**

1 Introduction . . . 49

2 Setting . . . 52

2.1 Notation, assumptions, and a partition of unity . . . 52

2.2 Examples of partitions of unity . . . 54

2.3 Poincaré–Friedrichs cut-off estimates . . . 54

2.4 An overlapping-patches estimate . . . 55

3 Localization of dual functional norms . . . 55

3.1 Context . . . 55

3.2 Examples of functionals $\mathcal{R}$ . . . 56

[1] This chapter is a preprint version of the article [*Jan Blechta, Josef Málek, and K.R. Rajagopal. On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion. 2019.*] submitted for publication in Society for Industrial and Applied Mathematics (SIAM) who is a copyright holder of the work. The preprint is separately available online at `https://arxiv.org/abs/1902.04853v1`.

[2] This chapter is a pre-copyedited, author-produced version of an article accepted for publication in *IMA Journal of Numerical Analysis* following peer review. The version of record [*Jan Blechta, Josef Málek, and Martin Vohralík. Localization of the $W^{-1,q}$ norm for local a posteriori efficiency. IMA J. Numer. Anal., 2019. Oxford University Press.*] is available online at `https://doi.org/10.1093/imanum/drz002`.

3.3 Motivation . . . 58
3.4 Main result . . . 59
3.5 Remarks . . . 61
4 Extensions . . . 62
4.1 Localization without the orthogonality condition . . . 62
4.2 Localization in vectorial setting . . . 67
5 Numerical illustration . . . 69
References . . . 77

III Analysis of PCD preconditioner for Navier-Stokes equations 81
1 Introduction . . . 81
2 Analysis of PCD in infinite-dimensional spaces . . . 84
2.1 Preliminaries . . . 84
2.2 A priori estimates and invertibility of the PCD operator . . . 89
2.3 GMRES iterations with the preconditioned saddle-point system . . . 95
2.4 Spectrum of the preconditioned Schur complement . . . 99
2.5 GMRES iterations with the preconditioned Schur complement . . . 101
2.6 PCD variants and boundary conditions . . . 101
2.7 Best-approximation property of the PCD correction . . . 103
3 Discretization of the preconditioner . . . 105
3.1 General considerations . . . 106
3.2 Case of continuous pressure discretizations . . . 111
3.3 Case of higher-order discontinuous pressure discretizations . . . 114
3.4 Case of piecewise constant pressure discretizations . . . 115
3.5 Historical remarks . . . 117
Appendix A Spectrum of bounded linear operators . . . 120
Appendix B Stability of contractive GMRES convergence under compact perturbations . . . 125
References . . . 130

Conclusion 135

List of publications 137

Preface

This work spans topics in mathematical and numerical analysis which are seemingly unrelated. The motivation comes from incompressible non-Newtonian fluid mechanics. Undoubtedly, understanding flows of non-Newtonian fluids is important for a broad range of applications in natural sciences and engineering and as such they attracted much of attention across a range of research fields, including, but not limited to, physics, scientific computing, and mathematics, specifically the theory of partial differential equations (PDEs) and numerical analysis. On the other hand, new tools and fundamental results are often designed and/or analyzed for only the simplest models (e.g., Newtonian fluids) in simple scenarios (e.g., no-slip boundary conditions). Hence there is space to develop successful techniques for more general complex problems. The three following chapters, which are possible to be read independently, present such an effort in thematically distant subjects, which share fluid mechanics as the greatest common divisor.

Chapter I presents a novel view on the classification of incompressible fluids. A previously unnoticed class of activated Euler fluids is discovered. A large-data existence theory of internal flows of this class of fluids is provided in a comprehensive combination of situations distinguishing steady and unsteady flows and a variety of boundary conditions.

Chapter II, motivated by its authors' work on a posteriori estimation techniques for non-Newtonian fluids [2], provides a theoretical result which is of independent interest in a posteriori error estimation theory. Specifically it is shown that the norm of functionals on dual Sobolev spaces $W^{-1,q}$, $1 \leq q \leq \infty$, is localizable by splitting into overlapping subdomains (e.g., mesh elements) assuming Galerkin orthogonality of the functional in question to the functions constituting a subordinate partition of unity. More general situations in which the Galerkin orthogonality is violated are discussed and a remedy is provided. The result is supported by several numerical experiments. The significance of the result is that it allows one to establish local efficiency of a posteriori error estimates for problems posed in $W^{-1,q}$ spaces. It was previously not obvious that this was possible unless $q = 2$, as the $W^{-1,q}$ norm, the natural residual norm for a wide class of non-linear PDE problems, does not have an obvious local structure.

Chapter III provides a novel theory for a modern preconditioning technique for incompressible fluids, the *pressure convection-diffusion* (PCD) preconditioner [6, section 9.2.1]. Although the chapter focuses on Navier-Stokes fluids, it is a potential step towards generalizing the technique to non-Newtonian fluids. The starting point of the work is the analysis of the preconditioner in appropriate infinite-dimensional spaces, an approach often coined *operator preconditioning*; see the survey monograph by Málek and Strakoš [8]. It is shown that the preconditioner and its inverse are under certain conditions well-defined. A priori estimates uniform in certain norms of data are provided and certain spectral properties of the preconditioned Schur complement are shown. Furthermore, GMRES convergence for the resulting preconditioned saddle-point system is shown to be almost contractive with a contraction factor given by the inf-sup constant of divergence and the norm of divergence. The analysis is stemming from a novel approach based on the observation that the preconditioned Schur complement is a compact perturbation of the Stokes Schur complement. This observation is the basis for the proposed methodology for the analysis of the quality of the preconditioner. The chapter continues by providing a new construction of the discrete variant of the preconditioner. It is further shown that some properties of the infinite-dimensional counterpart of the preconditioner are inherited by the discrete variant, e.g., surjectivity/injectivity and a priori norm/spectral bounds. We comment on why other results are rather difficult to transfer. The proposed discrete construction is then applied to several common finite element discretizations and novel variants of the preconditioner are derived while a previously published one is obtained as well.

The primary contribution of the chapter is that it provides very clear reasoning for (i) what boundary conditions should be used for construction of the preconditiner and (ii) how they should be incorporated in construction of the discrete preconditioner. It has been previously observed that (i) is a critical issue:

> These boundary conditions are not well understood, and a poor choice can critically affect performance. (Elman and Tuminaro [7, p. 257])

So far rather heuristic arguments were used to deal with (i). In our work the choice of boundary conditions in (i) emerges as a precondition to obtain the a priori estimates. In our opinion issue (ii) has so far been covered by rather confusing and contradictory accounts in the existing literature. Our results overcome this problem; we provide detailed comparison of our results and published accounts in Section III.3.5. Our analysis treats both variants of the PCD preconditioner given by different commutation orders in a unified way and provides a new argument for preference towards the variant due to Elman and Tuminaro [7].

Appendix III.B contains a new result concerning contractive convergence of the GMRES method. We consider a class of operators, which exhibit contractive GMRES convergence, i.e., the residual norms $\|r_k\|$ in subsequent GMRES steps fulfill

$$\frac{\|r_k\|}{\|r_{k-1}\|} \leq M \qquad \text{for all } k \in \mathbb{N}$$

with some $M < 1$ independent of the initial residual r_0. We show that when such operators are compactly perturbed, contractive convergence with the same factor is preserved asymptotically, i.e.,

$$\limsup_{k \to \infty} \frac{\|r_k\|}{\|r_{k-1}\|} \leq M.$$

Moreover, a measure of compactness, specifically inclusion of the perturbation in a p-Schatten class for some $p \geq 1$, determines the rate of the approach; precisely it holds

$$\left(\frac{\|r_k\|}{\|r_0\|} \right)^{\frac{1}{k}} \leq M + ck^{-\frac{1}{p}} \qquad \text{for all } k \in \mathbb{N}$$

with $c \geq 0$ independent of r_0.

Chapter I consists entirely of a preprint article by Blechta, Málek, and Rajagopal [1]. Chapter II is a reprint of a published article by Blechta, Málek, and Vohralík [3]. Chapter III is previously unpublished work solely by the present author. All chapters include a dedicated list of references. The numbering of equations, bibliographical references, theorems, figures, sections, etc. is local to each chapter, i.e., omits the respective chapter number, with the exception of a few cross-chapter references in Chapter III. In fact each chapter is self-contained and can be read independently.

The author would like to thank to his doctoral advisor Josef Málek, who always provides positive motivation. It was pleasure to work with him. Important acknowledgement belongs to the coauthors of Chapters I and II, K.R. Rajagopal and Martin Vohralík, whose patience can hardly be fully appreciated. The author thanks Howard Elman, Oliver Ernst, Martin Řehoř, and Zdeněk Strakoš for motivating the research in Chapter III through their inspiring work [5, 4, 8, 9, ...], for stimulating discussions, and for providing valuable feedback. A very special thanks for supporting my efforts belong to my family and to my partner Erin, who also carefully read Chapter III and provided grammar and stylistic corrections.

Jan Blechta, Prague, April 30, 2019

[1] J. Blechta, J. Málek, and K. Rajagopal. "On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion". Submitted. 2019. URL: https://arxiv.org/abs/1902.04853.

[2] J. Blechta, J. Málek, and M. Vohralík. "Generalized Stokes flows of implicitly constituted fluids: a posteriori error control and full adaptivity". In preparation. 2019.

[3] J. Blechta, J. Málek, and M. Vohralík. "Localization of the $W^{-1,q}$ norm for local a posteriori efficiency". In: *IMA J. Numer. Anal.* (Mar. 2019). DOI: 10.1093/imanum/drz002.

[4] M. Eiermann and O. G. Ernst. "Geometric aspects of the theory of Krylov subspace methods". In: *Acta Numer.* 10 (2001), pp. 251–312. DOI: 10.1017/S0962492901000046.

[5] H. C. Elman. "Preconditioning for the steady-state Navier-Stokes equations with low viscosity". In: *SIAM J. Sci. Comput.* 20.4 (1999), pp. 1299–1316. DOI: 10.1137/S1064827596312547.

[6] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics.* 2nd. Oxford University Press, 2014.

[7] H. C. Elman and R. S. Tuminaro. "Boundary Conditions in Approximate Commutator Preconditioners for the Navier-Stokes Equations". In: *Electronic Transactions on Numerical Analysis* 35 (2009), pp. 257–280. URL: http://etna.mcs.kent.edu/volumes/2001-2010/vol35/abstract.php?vol=35&pages=257-280.

[8] J. Málek and Z. Strakoš. *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs.* SIAM Spotlight Series. Philadelphia: SIAM, Jan. 2015. URL: http://bookstore.siam.org/sl01/.

[9] M. Řehoř. "Diffuse interface models in theory of interacting continua". 2018. URL: http://hdl.handle.net/20.500.11956/103641.

Chapter I

On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion[1]

1 Introduction

The concept of a fluid defies precise definition as one can always come up with a counter-example to that definition that seems to fit in with our understanding of what constitutes a fluid. As Goodstein [23] appropriately remarks "Precisely what do we mean by the term liquid? Asking what is a liquid is like asking what is life; we usually know when we see it, but the existence of some doubtful cases make it hard to define precisely." The concept of a fluid is treated as a primitive concept in mechanics, but unfortunately it does not meet the fundamental requirement of a primitive, that of being amenable to intuitive understanding. This makes the study under consideration that much more difficult as it is our intent to classify fluid bodies. In this study we shall consider a subclass of the idealization of a fluid, namely that of incompressible fluid bodies. While no material is truly incompressible, in many bodies the change of volume is sufficiently small to be ignorable. Our ambit will include at one extreme materials that could be viewed as incompressible Euler fluids and at the other extreme materials that offer so much resistance to flow that they are "rigid-like" in their response, with a whole host of "fluid-like" behavior exhibited by bodies whose response lie in between these two extremes, such as fluids exhibiting shear thinning/shear thickening, stress thinning/stress thickening, etc.

Before discussing the constitutive classification of fluid bodies, it would be useful to consider another type of classification that is used, namely that of flow classification with regard to the flows of a specific fluid, so that we do not confuse these two types of classifications. One of the most useful approximations and an integral part of fluid dynamics is the boundary layer approximation for the flow of a Navier-Stokes fluid (see Prandtl [69], Schlichting [80]). The main tenet of the approximation is the notion that for flows of a Navier-Stokes fluid past a solid boundary, at sufficiently high Reynolds number, the vorticity is confined to a thin region adjacent to the solid boundary. In this region, referred to as the boundary layer, the flow is dominated by the effects of viscosity while these effects fade away as one moves further away from the solid boundary. Sufficiently far from the boundary, the effects of viscosity are negligible and the equations governing the flow are identical to those for an Euler (ideal) fluid

[1] This chapter is a preprint version of the article [*Jan Blechta, Josef Málek, and K.R. Rajagopal. On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion. 2019.*] submitted for publication in Society for Industrial and Applied Mathematics (SIAM) who is a copyright holder of the work. The preprint is separately available online at `https://arxiv.org/abs/1902.04853v1`.

and one solves the problem by melding together the solution for the Euler fluid far from the boundary and the boundary layer approximation in a thin layer adjacent to the boundary. The reason for developing such an approach is the fact that in the "boundary layer region" an approximation is obtained for the Navier-Stokes equations that is more amenable to analysis than the fully non-linear equations. It is important to bear in mind that boundary layer theory is an approximation of the Navier-Stokes equations in different parts of the flow domain, that which is immediately adjacent to the solid boundary and that which is away from the solid boundary. Great achievements in the field of aerodynamics are a testimony to the efficacy and usefulness of such an approximation with regard to solving analytically or computationally relevant problems in a particular geometrical setting. On the other hand, rigorous analysis of the Prandtl boundary layer equations is, despite significant effort, far from being satisfactory (see [4], [34, 35, 36], [45, 46], [53], [54], [64], [67, 66], [79, 78]). An alternative viewpoint for modeling the boundary layer phenomena might thus bring some new insight on this issue.

The boundary layer approximation is not a constitutive approximation based on different flow regimes though it seems to resemble such an approximation. That is, one does not assume different constitutive assumptions for different regions in the flow domain, based on some kinematical or other criterion, but based on the value for the Reynolds number one merely carries out an approximation of the Navier-Stokes equation in the flow domain. It is possible, for instance, to assign different constitutive relations, based on the shear rate, namely the fluid being an Euler fluid below a certain shear rate and a Navier-Stokes or a non-Newtonian fluid above the critical shear rate (such a classification is considered in Section 2.5), or as another possibility a non-Newtonian fluid if the shear rate is below a certain value and a Navier-Stokes fluid above that shear rate, or any such assumption for the constitutive response of the material, and to solve the corresponding equations for the balance of linear momentum in the different flow domains. Such distinct constitutive responses below and above a certain kinematical criterion is akin to models for the inelastic response of bodies wherein below a certain value of the strain or stress, the body behaves as an elastic body while for values above the critical value the body responds in an inelastic manner, which in turn might lead to certain parts of a body to respond like an elastic body while other parts could be exhibiting inelastic response. To make matters clear, in a solid cylindrical body that is undergoing torsion, a yield condition based on the strain would lead to the body beyond a certain radius to respond inelastically while below that threshold for the radius it responds as an elastic body. In such an approach different constitutive relations are used in different domains while in the classical boundary layer theory one uses approximation of the equations of motion of a particular fluid.

In this paper, we adopt the approach of assuming different constitutive response relationships in different flow domains of the fluid, based on the value of the shear rate or the value of the shear stress. We consider the possibility that the character of the fluid changes when the certain "activation" criterion is reached. Here we consider an "activation" criterion that is based on the shear rate or the shear stress, but it could be any other criterion, say for instance the level of the electrical field in an electrorheological fluid, or the temperature which changes the character of the material from a fluid to a gas or a fluid to a solid, etc. Such an approach also provides an alternative way to viewing the classical boundary layer approximation in that it allows the fluid to behave like an Euler fluid in a certain flow domain and a Navier-Stokes fluid elsewhere. Furthermore, based on other criteria such as the Reynolds number we can carry out further approximations with regard to governing equations in the different flow domains.

Within the context of the Navier-Stokes theory, boundary layers occur at flows at sufficiently high Reynolds numbers. However, in the case of some non-Newtonian fluids it is possible to have regions that are juxtaposed to a solid boundary where the vorticity is concentrated even in the case of creeping flow, i.e., flows wherein the inertial effect is neglected when the Reynolds number is zero (see Mansutti and Rajagopal [62], [70], [38]). Thus, boundary layers are connected with the nonlinearities in the governing equation and are not a consequence of just high Reynolds numbers. Boundary layers can also occur at high Reynolds number in non-Newtonian fluids of the differential type (see Mansutti et al. [61]) and of the integral type (see Rajagopal and Wineman [74]). It is also possible that in non-Newtonian fluids one can have multiple decks with dominance of different physical mechanisms in the different decks, and in these different layers one can have the effects of viscosity, elasticity, etc., being significant, the delineation once again being determined within the context of a specific governing equation (see Rajagopal et al. [76, 75]). On the other hand, we could have a more complicated situation wherein the flow is characterized by different constitutive equations in different domains, and in these different domains it might happen that one can further delineate different subregions.

In the first part of this study, we provide a systematic classification of the response of incompressible fluid-like materials ranging from the ideal Euler fluid to non-Newtonian fluids that exhibit shear thinning/shear thickening, stress thinning/stress thickening, as well as those responses where the constitutive character of the material changes due to a threshold based on a kinematical, thermal, stress or some other quantity (an example of the same is the Bingham fluid which does not flow below a certain value of the shear stress and starts to flow once the threshold is overcome) based on some criterion concerning the level of shear rate or shear stress. We also provide a systematic study of both activated and non-activated boundary conditions ranging from free-slip to no-slip. In carrying out our classification, we come across the delineation of a class of fluids that, to our best knowledge, seems to have not been studied by fluid dynamicists. This class of fluids is characterized by the following intriguing dichotomy: (i) when the shear rate is below a certain critical value the fluid behaves as the Euler fluid (i.e., there is no effect of the viscosity, the shear stress vanishes), on the other hand (ii) if the shear rate exceeds the critical value, dissipation starts to take place and fluid can respond as a shear (or stress) thinning or thickening fluid or as a Navier-Stokes fluid. Implicit constitutive theory, cf. [72] and also [71, 73], provides an elegant framework to express such responses involving the activation criterion in a compact and elegant manner that is also more suitable for further mathematical and computational analysis.

In the second part of the paper, we study the mathematical properties of three-dimensional internal flows in bounded smooth domains for fluids belonging to this new class. We subject such flows to different types of boundary conditions including no-slip, Navier-slip, free-slip and activated boundary conditions like stick-slip. For this class of fluids and boundary conditions we prove the global-in-time existence of a weak solution in the sense of Leray to initial and boundary value problems.

2 Classification of incompressible fluids

Incompressible fluids are subject to the restriction on the admissible class of the velocity fields $\boldsymbol{v}$ of the form

$$\operatorname{div} \boldsymbol{v} = 0\,,$$

which can be written alternatively as

$$\operatorname{tr} \mathbb{D} = \mathbb{D} : \mathbb{I} = 0\,, \tag{2.1}$$

where $\mathbb{D}$ (sometimes denoted $\mathbb{D}\boldsymbol{v}$) stands for the symmetric part of the velocity gradient, i.e., $\mathbb{D} = \frac{1}{2}(\nabla \boldsymbol{v} + (\nabla \boldsymbol{v})^T)$, $\boldsymbol{v}$ being the velocity.

Due to this restriction, it is convenient to split the Cauchy stress tensor $\mathbb{T}$ into its traceless (deviatoric) part $\mathbb{S}$ and the mean normal stress, denoted by m (more frequently expressed as $-p$), i.e.,

$$\mathbb{S} = \mathbb{T} - \frac{1}{3}(\operatorname{tr}\mathbb{T})\mathbb{I} \quad \text{and} \quad -p = m = \frac{1}{3}\operatorname{tr}\mathbb{T}\,. \tag{2.2}$$

Hence

$$\mathbb{T} = m\mathbb{I} + \mathbb{S} = -p\mathbb{I} + \mathbb{S}\,,$$

and, in virtue of (2.1), the stress power $\mathbb{T} : \mathbb{D}$ satisfies

$$\mathbb{T} : \mathbb{D} = \mathbb{S} : \mathbb{D}\,.$$

The main result of this section will be the classification of fluids using a simple framework that is characterized by a relation between $\mathbb{S}$ and $\mathbb{D}$, i.e., we are interested in materials whose response can be incorporated into the setting given by the *implicit* constitutive equation

$$\mathcal{G}(\mathbb{S}, \mathbb{D}) = \mathbb{O}\,. \tag{2.3}$$

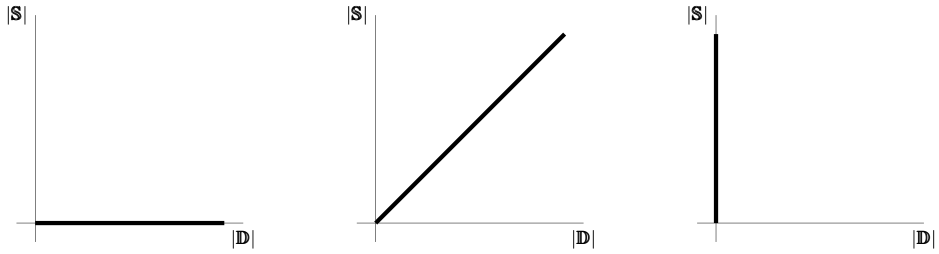


Figure 1: From left to right, response of the Euler fluid (2.6), the Navier-Stokes fluid (2.4) (or (2.5)), and fluid allowing only motions fulfilling (2.7)

The only restriction that we place is the requirement that response has to be *monotone.* For relevant discussion concerning non-monotone responses, we refer the reader to [57], [50], and [40].

The incompressible Navier-Stokes fluid is a special sub-class of (2.3) where the relation between $\mathbb{S}$ and $\mathbb{D}$ is linear. This can be written either as

$$\mathbb{S} = 2\nu_* \mathbb{D} \quad \text{with } \nu_* > 0\,, \tag{2.4}$$

where ν_* is called the (shear) *viscosity*, or as

$$\mathbb{D} = \alpha_* \mathbb{S} := \frac{1}{2\nu_*}\mathbb{S} \quad \text{with } \alpha_* > 0\,, \tag{2.5}$$

where the coefficient α_* is called the *fluidity.* Note that the stress power takes then the form

$$\mathbb{S} : \mathbb{D} = 2\nu_* |\mathbb{D}|^2 = \alpha_* |\mathbb{S}|^2\,.$$

There are two limiting cases when the stress power $\mathbb{S} : \mathbb{D}$ vanishes. Either

$$\mathbb{S} = \mathbb{O}, \tag{2.6}$$

which implies that the fluid under consideration is the incompressible Euler fluid ($\mathbb{T} = -p\mathbb{I}$, see (2.2)), or

$$\mathbb{D} = \mathbb{O} \quad \text{for all admissible flows.} \tag{2.7}$$

The latter corresponds to the situation where the body admits merely rigid body motions. More precisely, the flows fulfilling (2.7) can be characterized through

$$\boldsymbol{v}(t,x) = \boldsymbol{a}(t) \times x + \boldsymbol{b}(t) \quad \text{in all admissible flows.}$$

Response of models (2.4) (or (2.5)), (2.6), and (2.7) is shown in the Figure 1.

2.1 Classical power-law fluids

Classical power law fluids are described by

$$\mathbb{S} = 2\tilde{\nu}_* |\mathbb{D}|^{r-2}\mathbb{D}\,, \tag{2.8}$$

which leads to

$$\mathbb{S} : \mathbb{D} = 2\tilde{\nu}_* |\mathbb{D}|^r\,.$$

Since $|\mathbb{D}|^{r-2}\mathbb{D}$ should have meaning for $|\mathbb{D}| \to 0$, we require a lower bound on r, namely

$$r > 1\,. \tag{2.9}$$

Otherwise, if $r = 1$ then $\lim_{|\mathbb{D}|\to 0} \mathbb{D}/|\mathbb{D}|$ does not exist, and if $r < 1$ then $|\mathbb{S}| \to +\infty$ and stress concentration occurs at points where $\mathbb{D}$ vanishes (i.e., $\mathbb{S}$ plays the role of penalty for point where $\mathbb{D}$ could vanish).

In what follows we study power-law fluids with the power-law index satisfying (2.9) and we shall investigate the responses of these fluids for $r \to 1$ and $r \to \infty$. The latter corresponds to the case when the dual exponent $r' := r/(r-1)$ tends to 1.

We introduce the *generalized viscosity* through

$$\nu_{\mathrm{g}}(|\mathbb{D}|) = \tilde{\nu}_* |\mathbb{D}|^{r-2} . \tag{2.10}$$

In order to have the same units for ν_{g} as for the viscosity ν_* that appears in the formula for the Navier-Stokes fluid (see (2.4)), $\mathbb{D}$ should scale as d_* that has the unit s^{-1}. Thus, we replace (2.10) by

$$\nu_{\mathrm{g}}(|\mathbb{D}|) = \nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \qquad \text{where } [d_*] = \mathrm{s}^{-1} \text{ and } [\nu_*] = \mathrm{kg}\,\mathrm{m}^{-1}\,\mathrm{s}^{-1}$$

and we replace (2.8) by

$$\mathbb{S} = 2\nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \mathbb{D} \qquad \text{with } [d_*] = \mathrm{s}^{-1} \text{ and } [\nu_*] = \mathrm{kg}\,\mathrm{m}^{-1}\,\mathrm{s}^{-1}. \tag{2.11}$$

Of course, ν_* in (2.11) and ν_* in (2.4) are different in general; they however have the same units.

On considering (2.11), we notice the following equivalence[2]

$$\mathbb{S} = 2\nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \mathbb{D} \quad \Longleftrightarrow \quad \mathbb{D} = \frac{1}{2\nu_*} \left(\frac{|\mathbb{S}|}{2\nu_* d_*}\right)^{\frac{2-r}{r-1}} \mathbb{S} \,, \tag{2.12}$$

which gives rise to the following expressions for the generalized viscosity and *generalized fluidity*

$$\nu_{\mathrm{g}}(|\mathbb{D}|) = \nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \qquad \text{and} \qquad \alpha_{\mathrm{g}}(|\mathbb{S}|) = \frac{1}{2\nu_*} \left(\frac{|\mathbb{S}|}{2\nu_* d_*}\right)^{\frac{2-r}{r-1}} .$$

It also allows us to express the stress power in the form ($r' := r/(r-1)$)

$$\begin{aligned} \mathbb{S} : \mathbb{D} &= \left(\frac{1}{r} + \frac{1}{r'}\right) \mathbb{S} : \mathbb{D} = \frac{1}{r}\mathbb{S} : \mathbb{D} + \frac{1}{r'}\mathbb{S} : \mathbb{D} \\ &= 2\nu_* d_*^2 \left(\frac{1}{r}\left(\frac{|\mathbb{D}|}{d_*}\right)^{r} + \frac{1}{r'}\left(\frac{|\mathbb{S}|}{2\nu_* d_*}\right)^{r'}\right) . \end{aligned}$$

Summarizing,

$$\mathbb{S} = 2\nu_{\mathrm{g}}(|\mathbb{D}|^2)\mathbb{D} = 2\nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \mathbb{D} \iff \mathbb{D} = \alpha_{\mathrm{g}}(|\mathbb{S}|^2)\mathbb{S} = \frac{1}{2\nu_*} \left(\frac{|\mathbb{S}|}{2\nu_* d_*}\right)^{r'-2} \mathbb{S} \,, \tag{2.13}$$

emphasizing that the equivalence in (2.13) holds only if $r \in (1, +\infty)$ (which is equivalent to $r' \in (1, +\infty)$).

Generalizing the approach used in [60], we will investigate the limits of $\mathbb{S}$ and ν_{g} as $\mathbb{D}$ tends to zero or infinity, or vice versa, study limits of $\mathbb{D}$ and α_{g} as $\mathbb{S}$ vanishes or tends to infinity.

[2] Indeed, starting for example from the formula on the left-hand side of (2.12), we conclude that

$$|\mathbb{S}| = \frac{2\nu_*}{d_*^{r-2}} |\mathbb{D}|^{r-1},$$

which implies

$$|\mathbb{D}|^{2-r} = \left(\frac{d_*^{r-2}}{2\nu_*}|\mathbb{S}|\right)^{\frac{2-r}{r-1}} .$$

Hence

$$\mathbb{D} = \frac{1}{2\nu_*}\left(\frac{|\mathbb{D}|}{d_*}\right)^{2-r}\mathbb{S} = \left(\frac{1}{2\nu_*}\right)^{1+\frac{2-r}{r-1}} \left(d_*^{r-2}\right)^{\frac{2-r}{r-1}+1} |\mathbb{S}|^{\frac{2-r}{r-1}}\mathbb{S} = \frac{1}{2\nu_*}\left(\frac{|\mathbb{S}|}{2\nu_* d_*}\right)^{\frac{2-r}{r-1}}\mathbb{S},$$

which leads to the formula on the right-hand side of (2.12).

Letting $|\mathbb{D}| \to 0+$ we obtain, starting from the formula on the left-hand side of (2.13),

$$\begin{aligned}
&|\mathbb{S}| \to 0 && \text{if } r > 1, \\
&|\mathbb{S}| \le 2\nu_* d_* && \text{if } r = 1, \qquad (\text{as } |\mathbb{D}| \to 0+) \\
&|\mathbb{S}| \to +\infty && \text{if } r < 1,
\end{aligned}$$

and

$$\begin{aligned}
|\nu_{\mathrm{g}}(|\mathbb{D}|)| &\to 0 && \text{if } r > 2, \\
\nu_{\mathrm{g}}(|\mathbb{D}|) &= 2\nu_* && \text{if } r = 2, \qquad (\text{as } |\mathbb{D}| \to 0+) \\
|\nu_{\mathrm{g}}(|\mathbb{D}|)| &\to +\infty && \text{if } r < 2.
\end{aligned} \tag{2.14}$$

Thus, we note that the Cauchy stress $\mathbb{T}$ in the fluid tends to a purely spherical stress when r is greater than 1 and the norm of $\mathbb{D}$ tends to $0+$, or put differently, the constitutive relation for the fluid reduces to that for an Euler fluid.

Similarly, letting $|\mathbb{D}| \to +\infty$ we have

$$\begin{aligned}
&|\mathbb{S}| \to +\infty && \text{if } r > 1, \\
&|\mathbb{S}| \le 2\nu_* d_* && \text{if } r = 1, \qquad (\text{as } |\mathbb{D}| \to +\infty) \\
&|\mathbb{S}| \to 0 && \text{if } r < 1,
\end{aligned}$$

and

$$\begin{aligned}
|\nu_{\mathrm{g}}(|\mathbb{D}|)| &\to +\infty && \text{if } r > 2, \\
\nu_{\mathrm{g}}(|\mathbb{D}|) &= 2\nu_* && \text{if } r = 2, \qquad (\text{as } |\mathbb{D}| \to +\infty) \\
|\nu_{\mathrm{g}}(|\mathbb{D}|)| &\to 0 && \text{if } r < 2.
\end{aligned}$$

In order to investigate the behavior of $\mathbb{D}$ and fluidity in the limiting case, it is useful to employ the expression on the right-hand side of (2.13). Thus, for $|\mathbb{S}| \to 0+$, we get

$$\begin{aligned}
&|\mathbb{D}| \to 0 && \text{if } r' > 1, \\
&|\mathbb{D}| \le d_* && \text{if } r' = 1, \qquad (\text{as } |\mathbb{S}| \to 0+) \\
&|\mathbb{D}| \to +\infty && \text{if } r' < 1,
\end{aligned}$$

and

$$\begin{aligned}
|\alpha_{\mathrm{g}}(|\mathbb{S}|)| &\to 0 && \text{if } r' > 2, \\
\alpha_{\mathrm{g}}(|\mathbb{S}|) &= \frac{1}{2\nu_*} && \text{if } r' = 2, \qquad (\text{as } |\mathbb{S}| \to 0+) \\
|\alpha_{\mathrm{g}}(|\mathbb{S}|)| &\to +\infty && \text{if } r' < 2.
\end{aligned} \tag{2.15}$$

Similarly, letting $|\mathbb{S}| \to +\infty$ we have

$$\begin{aligned}
&|\mathbb{D}| \to +\infty && \text{if } r' > 1, \\
&|\mathbb{D}| \le d_* && \text{if } r' = 1, \qquad (\text{as } |\mathbb{S}| \to +\infty) \\
&|\mathbb{D}| \to 0 && \text{if } r' < 1,
\end{aligned}$$

and

$$\begin{aligned}
|\alpha_{\mathrm{g}}(|\mathbb{S}|)| &\to +\infty && \text{if } r' > 2, \\
\alpha_{\mathrm{g}}(|\mathbb{S}|) &= \frac{1}{2\nu_*} && \text{if } r' = 2, \qquad (\text{as } |\mathbb{S}| \to +\infty) \\
|\alpha_{\mathrm{g}}(|\mathbb{S}|)| &\to 0 && \text{if } r' < 2.
\end{aligned}$$

Next, we study the response of the classical power-law fluid with regard to its dependence on the value of power-law index ($r \to 1+$ and $r' \to 1+$). Figure 2 illustrates behavior for both large r and r' approaching 1.

Letting $r \to 1+$ in (2.13), we observe that, for $\mathbb{D} \neq \mathbb{O}$, $\mathbb{S} = 2\nu_* d_* \frac{\mathbb{D}}{|\mathbb{D}|}$ (and thus $|\mathbb{S}| \le 2\nu_* d_*$), while for $\mathbb{D} = \mathbb{O}$ and for any $\mathbb{A} \in \mathbb{R}^{3\times 3}_{sym}$ such that $|\mathbb{A}| \le 2\nu_* d_*$ we can find a sequence $\{\mathbb{D}_n\}_{n=1}^{\infty}$ converging to zero and

$$\lim_{n\to\infty} 2\nu_* d_* \frac{\mathbb{D}_n}{|\mathbb{D}_n|} = \mathbb{A}\,.$$

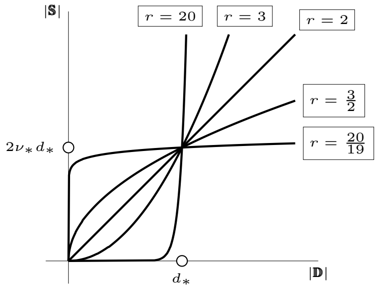


Figure 2: Response of the power-law model (2.11) for various values of r

Consequently, (2.13) for $r \to 1+$ approximates the response that could be referred to as *rigid/free-flow like behavior*:

$$\begin{aligned} &\text{if } |\mathbb{D}| \neq 0 \quad \text{then } \mathbb{S} = 2\nu_* d_* \frac{\mathbb{D}}{|\mathbb{D}|}, \\ &\text{if } |\mathbb{D}| = 0 \quad \text{then } |\mathbb{S}| \leq 2\nu_* d_*. \end{aligned} \tag{2.16}$$

Instead of viewing (2.16) as multivalued response (both in the variables $\mathbb{D}$ and $\mathbb{S}$), it is possible to write (2.16) as a continuous graph over the Cartesian product $\mathbb{R}^{3\times 3} \times \mathbb{R}^{3\times 3}$ (see the framework (2.3)) defined through a (scalar) equation

$$(|\mathbb{S}| - 2\nu_* d_*)^+ + \left|2\nu_* d_* \mathbb{D} - |\mathbb{D}|\mathbb{S}\right| = 0. \tag{2.17}$$

For determining the behavior of (2.13) as $r \to +\infty$ we prefer to study $(2.13)_2$ for $r' \to 1+$ and analogous to the above consideration we observe that

$$\begin{aligned} &\text{if } |\mathbb{S}| \neq 0 \quad \text{then } \mathbb{D} = \frac{d_*}{2\nu_*} \frac{\mathbb{S}}{|\mathbb{S}|}, \\ &\text{if } |\mathbb{S}| = 0 \quad \text{then } |\mathbb{D}| \leq \frac{d_*}{2\nu_*}. \end{aligned} \tag{2.18}$$

We can call this response *Euler/rigid like response*. We can again rewrite (2.18) as

$$(2\nu_* |\mathbb{D}| - d_*)^+ + \left|2\nu_* |\mathbb{S}|\mathbb{D} - d_* \mathbb{S}\right| = 0. \tag{2.19}$$

The models (2.16) and (2.18) are examples of fluids described within the context of an activation criterion. More examples will be discussed in Subsections 2.4 and 2.5. The slash formalism `name1/name2` (that we will use also below) means that material behaves as `name1` before activation and as `name2` after the activation criterion is met.

2.2 Generalized power-law fluids and stress power-law fluids

The formula (2.13) suggests the introduction of *generalized power-law fluids* and *generalized stress power-law fluids* by requiring that, for the former,

$$\mathbb{S} = 2\nu_g \left(|\mathbb{D}|^2\right) \mathbb{D}$$

and, for the latter,

$$\mathbb{D} = \alpha_g \left(|\mathbb{S}|^2\right) \mathbb{S}$$

where ν_g and α_g are non-negative continuous functions referred to as *the generalized viscosity* and *the generalized fluidity*. The quantities $|\mathbb{D}|^2 = \operatorname{tr} \mathbb{D}^2$ and $|\mathbb{S}|^2 = \operatorname{tr} \mathbb{S}^2$ representing the second invariants of $\mathbb{D}$ and $\mathbb{S}$, respectively, can be viewed as natural higher dimensional generalizations of the shear-rate and the shear stress, respectively.

We further introduce *zero shear rate viscosity* as

$$\nu_0 := \lim_{|\mathbb{D}| \to 0} \nu_g \left(|\mathbb{D}|^2\right)$$

and *zero shear stress fluidity* through

$$\alpha_0 := \lim_{|\mathbb{S}|\to 0} \alpha_{\mathrm{g}}\left(|\mathbb{S}|^2\right).$$

It follows from (2.14) that for classical power-law fluids the zero shear rate viscosity vanishes for $r > 2$, is finite for $r = 2$, and blows up if $r \in (1, 2)$. A similar behavior can be inferred from (2.15) for the zero shear stress fluidity: for $r' > 2$ (i.e., for $r \in (1, 2)$) α_0 is zero, for $r = 2$ it is positive, and for $r' \in (1, 2)$ (it means for $r > 2$) the generalized fluidity becomes singular in the vicinity of the origin.

Such behavior is not experimentally observed in any fluid; more frequently both the zero shear rate viscosity and zero shear stress fluidity are finite. The most popular generalizations of the classical power-law fluid that exhibit these features as $|\mathbb{D}| \to 0+$ (resp. $|\mathbb{S}| \to 0+$) take the form

$$\mathbb{S} = 2\nu_* \left(\frac{1}{2} + \frac{1}{2}\frac{|\mathbb{D}|^2}{d_*^2}\right)^{\frac{r-2}{2}} \mathbb{D} \tag{2.20}$$

and

$$\mathbb{D} = \alpha_* \left(\frac{1}{2} + \frac{1}{2}\frac{|\mathbb{S}|^2}{(2\nu_* d_*)^2}\right)^{\frac{r'-2}{2}} \mathbb{S}. \tag{2.21}$$

We refer the reader to [57] for further details; here we emphasize two observations. First, although both (2.20) and (2.21) are invertible for $r \in (1, +\infty)$, (2.21) is not an inverse of (2.20) and vice versa (compare it to (2.13)). Second, both formulas are defined for all $r \in (-\infty, +\infty)$ and $r' \in (-\infty, +\infty)$ respectively. The relationship $r' = \frac{r}{r-1}$ is however understood as a relation valid for $r > 1$.

If $r < 1$, then $\mathbb{S}$ considered as a function of $\mathbb{D}$ given by (2.20) is not monotone; the same holds for (2.21) if $r' < 1$. We refer the reader to [57] for more details and to [50] for further nontrivial extensions.

2.3 Fluids that can be viewed as a mixture of power-law fluids

It is natural to consider the possibility that the total response of the fluid-like material is given as the sum of particular responses (a simplified scenario for the mixtures) of the individual (in our case two) contributors, i.e.,

$$\mathbb{S} = \mathbb{S}_1 + \mathbb{S}_2. \tag{2.22}$$

One may think of putting two dashpots into a parallel arrangement; the response of one dashpot captures behavior of a fluid A, and the response of another dashpot corresponds to a fluid B; see Figure 3. To illustrate the potential of this setting, we consider for illustration three examples:

(i) $\mathbb{S}_1 = 2\nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \mathbb{D}$ and $\mathbb{S}_2 = 2\tilde{\nu}_* \left(\frac{1}{2} + \frac{1}{2}\frac{|\mathbb{D}|^2}{d_*^2}\right)^{\frac{q-2}{2}} \mathbb{D}$ with $r \in (1, 2)$, $q > 2$;

(ii) $\mathbb{S}_1 = 2\nu_*\mathbb{D}$ and $\mathbb{S}_2$ fulfills $\mathbb{D} = \frac{1}{2\tilde{\nu}_*} \left(\frac{1}{2} + \frac{1}{2}\frac{|\mathbb{S}_2|^2}{(2\tilde{\nu}_* d_*)^2}\right)^{\frac{r'-2}{2}} \mathbb{S}_2$ with $r' \in (1, +\infty) \setminus \{2\}$;

(iii) $\mathbb{S}_1$ responds as in (2.17) and $\mathbb{S}_2$ responds as in (2.19).

The responses of fluids modeled by the constitutive expressions (i) and (ii) are shown in Figure 4 for some choice of parameters. Further examples will be provided in Subsection 2.5.

2.4 Fluids with bounded shear rate or bounded shear stress

We have seen in Subsection 2.1 that the models (2.16) and (2.18) exhibit an interesting feature, namely, the stress $\mathbb{S}$ is bounded as the symmetric part of the velocity gradient is varied, and vice versa. While there are several mathematical advantages to the stress $\mathbb{S}$ expressed as a function of $\mathbb{D}$, as this would greatly simplify the number of equations that one needs to consider when one substitutes it into the balance of linear momentum (even an explicit expression of the symmetric part of the velocity gradient as a function of $\mathbb{S}$ might lead to a simplified structure

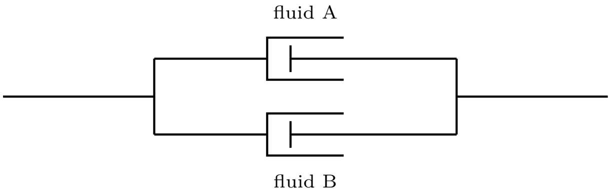


Figure 3: 1D mechanical analogue of an additive stress model (2.22)

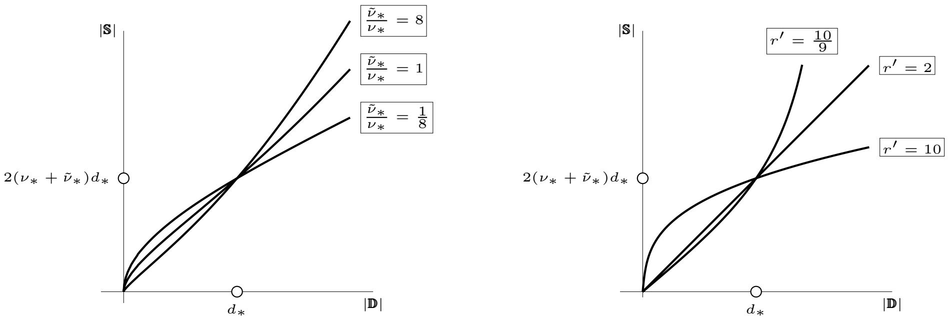


Figure 4: Response of models represented by (2.22). On the left, the case (i) with $r = \frac{3}{2}$ and $q = \frac{5}{2}$. On the right, the case (ii) with $\nu_* = \tilde{\nu}_*$.

for the equations) experiments on colloidal fluids clearly show that a fully implicit theory is necessary (see for example [12, 39] and further references in [68]). When considering the models (2.20) and (2.21) with $r = 1$ and $r' = 1$ (and adjusting a scaling by factor $\sqrt{2}$), we obtain

$$\mathbb{S} = 2\nu_* \frac{\mathbb{D}}{\sqrt{1 + \frac{|\mathbb{D}|^2}{d_*^2}}} \tag{2.23}$$

and

$$\mathbb{D} = \alpha_* \frac{\mathbb{S}}{\sqrt{1 + \frac{|\mathbb{S}|^2}{(2\nu_* d_*)^2}}}, \tag{2.24}$$

and we observe that, in the case of (2.23),

$$|\mathbb{S}| \leq 2\nu_* d_* \quad \text{for all } \mathbb{D},$$

and, in the case of (2.24),

$$|\mathbb{D}| \leq d_* \quad \text{for all } \mathbb{S}.$$

It is convenient to generalize (2.23) and (2.24) in the following manner: for parameters $a, b \in (0, +\infty)$ consider

$$\mathbb{S} = 2\nu_* \frac{\mathbb{D}}{\left(1 + \left(\frac{|\mathbb{D}|}{d_*}\right)^a\right)^{\frac{1}{a}}} \tag{2.25}$$

and

$$\mathbb{D} = \alpha_* \frac{\mathbb{S}}{\left(1 + \left(\frac{|\mathbb{S}|}{2\nu_* d_*}\right)^b\right)^{\frac{1}{b}}}. \tag{2.26}$$

In both cases, it is worth studying the behavior of the fluids for large a and b. When $a \to +\infty$ in (2.25), the constitutive relation approximates the response of the activated fluid which behaves as the Euler fluid prior to the activation and the magnitude of the stress remains bounded; analogously, when $b \to +\infty$ in (2.26), the constitutive relation approximates the response such that the magnitude of $\mathbb{D}$ remains bounded and the body admits merely rigid body motions till the activation takes place, see Figure 5.

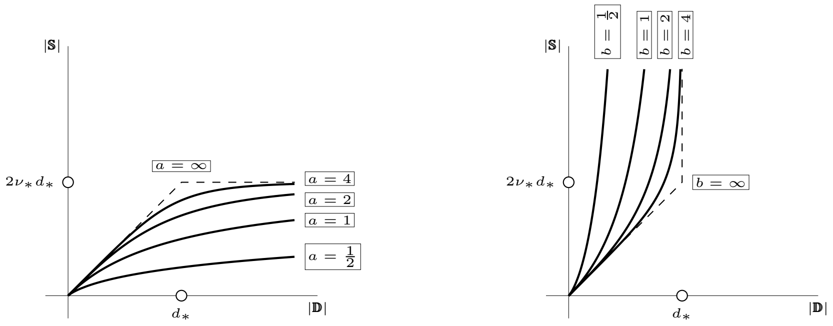


Figure 5: Response of stress-limiting model (2.25) on the left and shear-limiting model (2.26) on the right

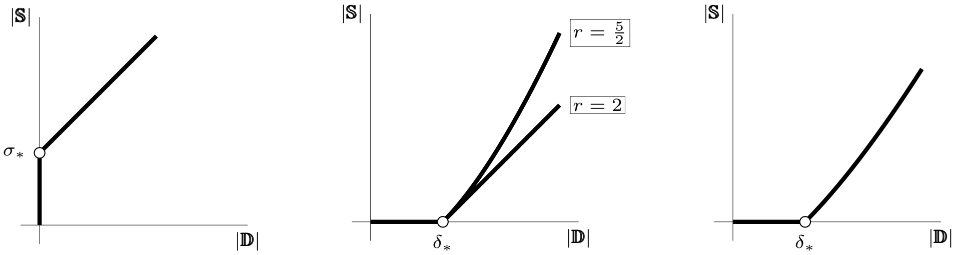


Figure 6: Response of the Bingham fluid (on the left), activated Euler/Navier-Stokes fluid (in the middle, $r = 2$), Euler/power-law fluid (in the middle, $r = \frac{5}{2}$, $d_* = \delta_*$), and Euler/Ladyzhenskaya fluid (on the right, $r = \frac{5}{2}$, $d_* = \delta_*$, $\nu_* = \tilde{\nu}_*$)

2.5 Activated fluids

In this section we study two classes of fluids: the first class is activation based on the value of the stress (similar in character to a Bingham fluid) while the second class is activation based on the value of the shear rate.

The first class of fluids that are studied flow only if the generalized shear stress $|\mathbb{S}| = \left(\operatorname{tr}\mathbb{S}^2\right)^{\frac{1}{2}}$ exceeds a certain critical value σ_*, referred to as *the yield stress.* Once the fluid flows, we assume the fluid behavior is described by the constitutive expression for a generalized power-law or a generalized stress power-law fluid. In the parts of the subdomain where $|\mathbb{S}|$ is below σ_* the fluid can only translate or rotate as a rigid body. Such responses are traditionally described (see [28]) through the dichotomy

$$\begin{aligned} |\mathbb{S}| \le \sigma_* \quad &\iff \quad \mathbb{D} = \mathbb{O}, \\ |\mathbb{S}| > \sigma_* \quad &\iff \quad \mathbb{S} = \sigma_* \frac{\mathbb{D}}{|\mathbb{D}|} + \mathbb{S}_2 \quad \text{with} \begin{cases} \text{either} & \mathbb{S}_2 = 2\nu_{\mathrm{g}}\left(|\mathbb{D}|^2\right)\mathbb{D}, \\ \text{or} & \mathbb{D} = \alpha_{\mathrm{g}}\left(|\mathbb{S}_2|^2\right)\mathbb{S}_2. \end{cases} \end{aligned} \tag{2.27}$$

In the case of the stress $\mathbb{S}_2 = 2\nu_*\mathbb{D}$ we obtain the constitutive representation for the *Bingham fluid* (see Figure 6 on the left) and if $\mathbb{S}_2 = 2\nu_{\mathrm{g}}\left(|\mathbb{D}|^2\right)\mathbb{D}$ then we obtain the constitutive representation for the *Herschel-Bulkley* fluid. It is worth of observing that (2.27) can be equivalently written within the context of the framework for implicit constitutive equations (2.3). Specifically, considering (2.27) with the expression $\mathbb{S}_2 = 2\nu_{\mathrm{g}}\left(|\mathbb{D}|^2\right)\mathbb{D}$ the equivalent formulation can be expressed as

$$2\nu_{\mathrm{g}}\left(|\mathbb{D}|^2\right)\mathbb{D} = \frac{\left(|\mathbb{S}| - \sigma_*\right)^+}{|\mathbb{S}|}\mathbb{S}$$

where $(t)^+ = \max\{t, 0\}$ for $t \in \mathbb{R}$.

On the other hand, considering (2.27) with the expression $\mathbb{D} = \alpha_{\mathrm{g}}\left(|\mathbb{S}_2|^2\right)\mathbb{S}_2$, the equivalent representation reads

$$\mathbb{D} = \alpha_{\mathrm{g}}\left(\left|\mathbb{S} - \sigma_*\frac{\mathbb{D}}{|\mathbb{D}|}\right|^2\right)\frac{\left(|\mathbb{S}| - \sigma_*\right)^+}{|\mathbb{S}|}\mathbb{S},$$

which is in our opinion worthy of detailed investigation.

The next class is a dual to (2.27) in the following sense. If the generalized shear rate $|\mathbb{D}|$ is below a critical value δ_*, the flow is frictionless. Inner friction between the fluid layers becomes important only when $|\mathbb{D}|$ exceeds δ_*. Then the fluid can flow as a Navier-Stokes fluid, or a power-law fluid (see Figure 6 on the right), or a generalized power-law fluid, or a generalized stress power-law fluid. To summarize, analogous to (2.27), we can describe such a response through the relation

$$\begin{aligned} |\mathbb{D}| \le \delta_* \quad &\iff \quad \mathbb{S} = \mathbb{O}, \\ |\mathbb{D}| > \delta_* \quad &\iff \quad \mathbb{D} = \delta_* \frac{\mathbb{S}}{|\mathbb{S}|} + \alpha_{\mathrm{g}}\left(|\mathbb{S}|^2\right)\mathbb{S}. \end{aligned}$$

It is not surprising that this relation can be written as an explicit relation for the stress $\mathbb{S}$ in terms of the symmetric part of the velocity gradient $\mathbb{D}$, namely

$$\alpha_{\mathrm{g}}\left(|\mathbb{S}|^2\right)\mathbb{S} = \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|}\mathbb{D}. \tag{2.28}$$

If there is ν_{g} such that

$$\mathbb{D} = \alpha_{\mathrm{g}}\left(|\mathbb{S}|^2\right)\mathbb{S} \quad \iff \quad \mathbb{S} = 2\nu_{\mathrm{g}}\left(|\mathbb{D}|^2\right)\mathbb{D} \tag{2.29}$$

then (2.28) can be written in the form ((2.29) describes the behavior of the fluid after activation)

$$\mathbb{S} = 2\nu_{\mathrm{g}}\left(|\mathbb{D}|^2\right)\frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|}\mathbb{D}.$$

Explicit equivalence holds for the Navier-Stokes fluid (see (2.5)) and for the standard power-law fluids (see (2.13)). Then we obtain the response

$$\mathbb{S} = 2\nu_* \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|}\mathbb{D} \tag{2.30}$$

and

$$\mathbb{S} = 2\nu_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2} \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|}\mathbb{D}, \tag{2.31}$$

respectively (see Figure 6 on the right). We call the response (2.30) *the Euler/Navier-Stokes fluid* and the response (2.31) *the Euler/power-law fluid.* If

$$\mathbb{S} = \left(2\nu_* + 2\tilde{\nu}_* \left(\frac{|\mathbb{D}|}{d_*}\right)^{r-2}\right) \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|}\mathbb{D},$$

with $r > 2$ we call this response *Euler/Ladyzhenskaya fluid*, since O. A. Ladyzhenskaya was the first to consider the generalization of the Navier-Stokes constitutive equation to the form

$$\mathbb{S} = \left(2\nu_* + 2\tilde{\nu}_*|\mathbb{D}|^{r-2}\right)\mathbb{D} \tag{2.32}$$

and showed that unsteady internal flows of such fluids in a bounded smooth container admit *unique* weak solutions if $r > \frac{5}{2}$ (or $\frac{d+2}{2}$ in general dimension d); see [20, 18] for improvement of the uniqueness result to $r \ge \frac{11}{5}$. Ladyzhenskaya used kinetic theory arguments to derive (2.32) with $r = 4$; see [48, 49, 47].

Simple shear flows of the Euler/Navier-Stokes fluid For the sake of illustration of response of the fluid (2.30) we consider a simple shear flow of such a fluid. In order to characterize such fluids it is also useful to consider a modified model which we call *the regularized Euler/Navier-Stokes fluid* given by

$$\mathbb{S} = 2\nu_* \left(\epsilon_* + \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|}\right)\mathbb{D} \tag{2.33}$$

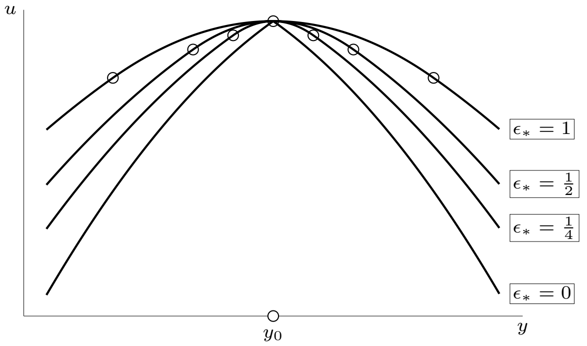


Figure 7: Simple shear flows of the regularized Euler/Navier-Stokes fluid for various values of the added viscosity $\epsilon_*\nu_*$ and fixed $C > 0$. Circles mark the point of activation $y = y_0 \pm \frac{\sqrt{2}\delta_*\epsilon_*}{2|C|}$ where $|u'| = \sqrt{2}\delta_*$. The degenerate case $\epsilon_* = 0$ is activated everywhere, i.e., $|u'| \geq \sqrt{2}\delta_*$. Note that the velocity profiles are determined up to an additive constant (because no boundary conditions are enforced); here we take $u_0 = 0$.

with an extra parameter $\epsilon_* \geq 0$. We will consider solutions of the balance equations (see (2.40), (2.41) below) for the response of the fluids described by (2.33) (and the degenerate case (2.30)) in $\mathbb{R}^2$, with the velocity taking the form

$$\boldsymbol{v}(x,y) = (u(y), 0) \qquad x, y \in \mathbb{R}$$

for some $u : \mathbb{R} \to \mathbb{R}$ absolutely continuous on every compact interval in $\mathbb{R}$. We note that now we deal with solutions of the governing equations in $\mathbb{R}^2$ so there are no boundary conditions involved. It is easy to check that if $\epsilon_* > 0$ then all such solutions of (2.40), (2.41), and (2.33) fulfill

$$\left(\epsilon_* + \mathcal{H}(|u'| - \sqrt{2}\delta_*)\right) u'' = -2C \quad \text{a.e. in } \mathbb{R}, \tag{2.34a}$$

$$p(x) = -2\nu_* Cx + p_0, \tag{2.34b}$$

with some $C \in \mathbb{R}$, $\mathcal{H}(t) = 1$ if $t > 0$ and $\mathcal{H}(t) = 0$ otherwise. The formulas (2.34) represent a generalization of the well-known equations for simple shear flows of the Navier-Stokes fluid, which is a special case with $\delta_* = 0$. In fact all the solutions of (2.34) take the form

$$u(y) = \begin{cases} -\frac{C}{\epsilon_*}(y-y_0)^2 + u_0 & |y-y_0| \leq \frac{\sqrt{2}\delta_*\epsilon_*}{2|C|}, \\ -\frac{C}{1+\epsilon_*}\left((y-y_0)^2 + \sqrt{2}\delta_*\left|\frac{y-y_0}{C}\right| - \epsilon_*\left(\frac{\sqrt{2}\delta_*}{2C}\right)^2\right) + u_0 & |y-y_0| \geq \frac{\sqrt{2}\delta_*\epsilon_*}{2|C|}, \end{cases} \tag{2.35a}$$

$$p(x) = -2\nu_* Cx + p_0, \tag{2.35b}$$

with any $C, y_0, u_0, p_0 \in \mathbb{R}$. In the interval $\left\{|y-y_0| \leq \frac{\sqrt{2}\delta_*\epsilon_*}{2|C|}\right\}$ the fluid is in the regime below the "activation" threshold (where $|u'| \leq \sqrt{2}\delta_*$) with the viscosity $\epsilon_*\nu_*$ while outside this interval the threshold is exceeded and the generalized viscosity has the value $\nu_*\left(\epsilon_* + 1 - \frac{\sqrt{2}\delta_*}{|u'|}\right)$. Taking the limit $\epsilon_* \to 0+$ one obtains

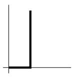
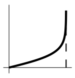
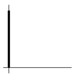
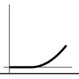
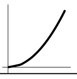
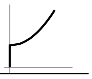
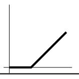
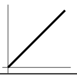
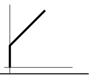
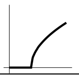
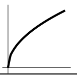
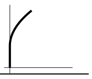
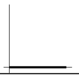
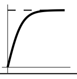
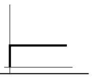
$$u(y) = -C\left((y-y_0)^2 + \sqrt{2}\delta_*\left|\frac{y-y_0}{C}\right|\right) + u_0, \tag{2.36a}$$

$$p(x) = -2\nu_* Cx + p_0, \tag{2.36b}$$

which indeed is a solution of the balance equations for the Euler/Navier-Stokes fluid (2.30) with any $C, y_0, u_0, p_0 \in \mathbb{R}$. Now the flow exceeds the activation threshold everywhere except of $y = y_0$ where $u'(y_0\pm) = \mp\sqrt{2}\delta_*\frac{C}{|C|}$ and the shear rate jumps there by virtue of the vanishing viscosity. In Figure 7 we display a family of solutions (2.35a) for varying ϵ_* and (2.36a) (matching $\epsilon_* = 0$) for fixed values of C, y_0.

Apart of the family of solutions (2.36) the Euler/Navier-Stokes fluid (2.30) admits also the simple shear flows which do not exceed the threshold $|u'| = \sqrt{2}\delta_*$, the viscosity is zero and

Table 1: Summary of systematic classification of fluid-like response with the corresponding $|\mathbb{D}|$-$|\mathbb{S}|$ diagrams

Euler/rigid		Navier-Stokes/limiting shear-rate		fluid body allowed to move only rigidly	
Euler/shear-thickening		shear-thickening		rigid/shear-thickening	
Euler/Navier-Stokes		Navier-Stokes		rigid/Navier-Stokes	
Euler/shear-thinning		shear-thinning		rigid/shear-thinning	
Euler		limiting		rigid/free-flow	
$|\mathbb{D}| \le \delta_* \iff \mathbb{S} = \mathbb{O}$		no activation		$|\mathbb{S}| \le \sigma_* \iff \mathbb{D} = \mathbb{O}$	

therefore any admissible velocity profile is a solution. Such solutions are characterized by

$$|u'| \le \sqrt{2}\delta_* \quad \text{a.e. in } \mathbb{R}, \tag{2.37a}$$

$$p = p_0, \tag{2.37b}$$

with some locally Lipschitz continuous $u : \mathbb{R} \to \mathbb{R}$ and $p_0 \in \mathbb{R}$. In fact the families (2.36) and (2.37) are all possible weak solutions for a simple shear flow of the Euler/Navier-Stokes fluid.

2.6 Classification of incompressible fluids

The previous exposition should indicate the broad spectrum of fluid responses that can be described within the setting

$$b\left(|\mathbb{D}|^2\right)\mathbb{D} = a\left(|\mathbb{S}|^2\right)\mathbb{S}, \tag{2.38}$$

where a and b are continuous (not necessarily always differentiable) functions.

The following Table 1 summarizes these observations in a different way, paying attention to the broad range of models covered by (2.38). It includes the Euler (frictionless) fluid at one extreme and a fluid that only moves rigidly at the other extreme and contains the responses ranging from the fluids enforcing the activation criterion $|\mathbb{D}| \le \delta_* \iff \mathbb{S} = \mathbb{O}$ through non-activated fluids to the fluids that are governed by the activation criterion $|\mathbb{S}| \le \sigma_* \iff \mathbb{D} = \mathbb{O}$. Vertically, the range of r is iterated from top to bottom: $r = +\infty$, $r \in (2, +\infty)$, $r = 2$, $r \in (1, 2)$, and $r = 1$. Thus, at the bottom left corner we have perfectly frictionless Euler fluid and at the top right corner we have a fluid that can only undergo rigid motions. In the middle of the table the Navier-Stokes model is placed.

2.7 Activated boundary conditions

Boundary conditions can have as much impact on the nature of the flow as the constitutive equation for the fluid in the bulk. We illustrate it explicitly in this subsection. Here, for the sake of clarity, we restrict ourselves to internal flows, i.e., we assume that

$$\boldsymbol{v} \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{2.39}$$

where $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^d$ denotes the mapping that assigns the outward unit normal vector to any $\boldsymbol{x} \in \partial\Omega$. The behavior of the fluid at the tangential direction near the boundary is described by the equations that reflect mutual interaction between the solid boundary and the fluid flowing adjacent to the boundary. These are constitutive equations which we shall concern ourselves with. In order to specify them, in the spirit of previous parts of the paper, we first recall some energy estimates.

Considering the constraint that the fluid can undergo only isochoric motions

$$\operatorname{div} \boldsymbol{v} = 0 \quad \text{in } \Omega, \tag{2.40}$$

and assuming, for simplicity, that the density is uniform, i.e., $\rho \equiv \rho_* > 0$, motions of such a fluid are described by the balance equations for linear and angular momenta that take the form (see also (2.2))

$$\begin{aligned} \rho_* \left(\frac{\partial \boldsymbol{v}}{\partial t} + \sum_{k=1}^{d} v_k \frac{\partial \boldsymbol{v}}{\partial x_k} \right) &= \operatorname{div} \mathbb{S} - \nabla p \qquad &&\text{in } \Omega, \\ \mathbb{S} &= \mathbb{S}^{\top} &&\text{in } \Omega. \end{aligned} \tag{2.41}$$

Forming the scalar product of the first equation with $\boldsymbol{v}$ and integrating the result over Ω, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \rho_* \frac{|\boldsymbol{v}|^2}{2} \mathrm{d}\boldsymbol{x} + \int_{\Omega} \operatorname{div} \left(\rho_* \frac{|\boldsymbol{v}|^2}{2} \boldsymbol{v} \right) \mathrm{d}\boldsymbol{x} = \int_{\Omega} \operatorname{div} (\mathbb{S}\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} - \int_{\Omega} \mathbb{S} : \mathbb{D} \, \mathrm{d}\boldsymbol{x} - \int_{\Omega} \operatorname{div} (p\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x},$$

where we have used (2.40) twice. Gauss' theorem and the requirement (2.39) then lead to

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \rho_* \frac{|\boldsymbol{v}|^2}{2} \mathrm{d}\boldsymbol{x} + \int_{\Omega} \mathbb{S} : \mathbb{D} \, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} (-\mathbb{S}) : (\boldsymbol{v} \otimes \boldsymbol{n}) \, \mathrm{d}S = 0, \tag{2.42}$$

where $\boldsymbol{a} \otimes \boldsymbol{b}$ denotes the second order tensor with the components $(\boldsymbol{a} \otimes \boldsymbol{b})_{ij} = a_i b_j$. In virtue of the symmetry of $\mathbb{S}$, see (2.41), we obtain

$$(-\mathbb{S}) : (\boldsymbol{v} \otimes \boldsymbol{n}) = (-\mathbb{S}) : (\boldsymbol{n} \otimes \boldsymbol{v}) = (-\mathbb{S}\boldsymbol{n})_{\boldsymbol{\tau}} \cdot \boldsymbol{v}_{\boldsymbol{\tau}},$$

where $\boldsymbol{z}_{\boldsymbol{\tau}}$ denotes the projection of $\boldsymbol{z} : \partial\Omega \to \mathbb{R}^d$ to the plane tangent to $\partial\Omega$ (at the point of $\partial\Omega$ under consideration). Finally, introducing the notation

$$\boldsymbol{s} := (-\mathbb{S}\boldsymbol{n})_{\boldsymbol{\tau}} \quad \text{(projection of the normal traction to the tangent plane)},$$

we can rewrite (2.42) in the form

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \rho_* \frac{|\boldsymbol{v}|^2}{2} \mathrm{d}\boldsymbol{x} + \int_{\Omega} \mathbb{S} : \mathbb{D} \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \boldsymbol{s} \cdot \boldsymbol{v}_{\boldsymbol{\tau}} \mathrm{d}S = 0.$$

The discussion in Section 2 thus far has been focused on discussing models within the context of the framework $\mathcal{G}(\mathbb{S}, \mathbb{D}) = \mathbb{O}$ (see (2.3)). In fact, the discussion concerned the restricted class of models of the form

$$a\left(|\mathbb{D}|^2\right) \mathbb{D} = b\left(|\mathbb{S}|^2\right) \mathbb{S}, \tag{2.43}$$

where a and b were non-negative continuous (not necessarily everywhere differentiable) functions.

In a manner similar to the class of models defined through (2.43), we could develop analogously the identical framework of relations linking $\boldsymbol{s}$ and $\boldsymbol{v}_{\boldsymbol{\tau}}$, i.e., to consider various classes of boundary conditions that fit the form

$$\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) = \boldsymbol{0},$$

where we deal with a (monotone) continuous function $\boldsymbol{h} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, or a more restrictive class

$$\tilde{a}\left(|\boldsymbol{v}_{\boldsymbol{\tau}}|^2\right) \boldsymbol{v}_{\boldsymbol{\tau}} = \tilde{b}\left(|\boldsymbol{s}|^2\right) \boldsymbol{s}, \tag{2.44}$$

where $\tilde{a}, \tilde{b} : [0, +\infty) \to [0, +\infty)$ are non-negative continuous functions.

We shall not consider the problem within the context of such generality for the following two reasons: (i) we do not have enough experimental data that would support nonlinear relations between $\boldsymbol{s}$ and $\boldsymbol{v}_{\boldsymbol{\tau}}$, and (ii) the extension of the framework developed to take into account such nonlinear relations is straightforward and follows in a manner similar to that discussed in Subsections 2.1–2.5.

In what follows, we restrict ourselves to both activated and non-activated responses that are (after activation) linear. In a manner similar to that in the introductory part of Section 2 where we considered the product $\mathbb{S}:\mathbb{D}$, we notice here that the product $\boldsymbol{s}\cdot\boldsymbol{v}_{\boldsymbol{\tau}}$ vanishes if either

$$\boldsymbol{s}=\boldsymbol{0} \quad \text{on } \partial\Omega, \tag{2.45}$$

or

$$\boldsymbol{v}_{\boldsymbol{\tau}}=\boldsymbol{0} \quad \text{on } \partial\Omega \text{ for all admissible flows.} \tag{2.46}$$

The condition (2.45) is referred to as the *free-slip* condition condition, expressing the fact that the boundary exhibits no friction to the flow, in the sense that the shear stress vanishes, and fluid flows tangentially to the boundary. On the other hand, the condition (2.46), referred to as the *no-slip* boundary condition, requires that flows adhere to the boundary. It has the character of a boundary constraint.

A linear relation between $\boldsymbol{s}$ and $\boldsymbol{v}_{\boldsymbol{\tau}}$ is known as the *Navier-slip*:

$$\boldsymbol{s}=\gamma_*\boldsymbol{v}_{\boldsymbol{\tau}} \qquad \text{with } \gamma_*>0. \tag{2.47}$$

We consider two types of activated boundary conditions. First, the relation

$$\begin{aligned} |\boldsymbol{s}|\le s_* \quad &\Longleftrightarrow \quad \boldsymbol{v}_{\boldsymbol{\tau}}=\boldsymbol{0},\\ |\boldsymbol{s}|> s_* \quad &\Longleftrightarrow \quad \boldsymbol{s}=s_*\frac{\boldsymbol{v}_{\boldsymbol{\tau}}}{|\boldsymbol{v}_{\boldsymbol{\tau}}|}+\gamma_*\boldsymbol{v}_{\boldsymbol{\tau}}, \end{aligned} \tag{2.48}$$

which has been coined as the stick/slip condition in the literature, but which we will refer to as the *no-slip/Navier-slip* condition for consistency; s_* is the yield stress that is positive. It can be rewritten into the form (2.44) through its equivalent characterization

$$\gamma_*\boldsymbol{v}_{\boldsymbol{\tau}}=\frac{(|\boldsymbol{s}|-s_*)^+}{|\boldsymbol{s}|}\boldsymbol{s}. \tag{2.49}$$

In analogy, the second type of activated condition is given through the description

$$\begin{aligned} |\boldsymbol{v}_{\boldsymbol{\tau}}|\le v_* \quad &\Longleftrightarrow \quad \boldsymbol{s}=\boldsymbol{0},\\ |\boldsymbol{v}_{\boldsymbol{\tau}}|> v_* \quad &\Longleftrightarrow \quad \boldsymbol{v}_{\boldsymbol{\tau}}=v_*\frac{\boldsymbol{s}}{|\boldsymbol{s}|}+\frac{1}{\gamma_*}\boldsymbol{s}, \end{aligned} \tag{2.50}$$

where $v_*>0$. The equivalent description of (2.50), that can be referred to as the *free-slip/Navier-slip* condition, takes the form

$$\boldsymbol{s}=\gamma_*\frac{(|\boldsymbol{v}_{\boldsymbol{\tau}}|-v_*)^+}{|\boldsymbol{v}_{\boldsymbol{\tau}}|}\boldsymbol{v}_{\boldsymbol{\tau}}. \tag{2.51}$$

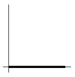
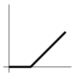
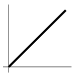
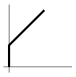
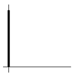
These responses are summarized in the Table 2.

Simple shear flows of the Navier-Stokes fluid and the Euler/Navier-Stokes fluid subject to activated boundary conditions Finally, in order to emphasize the role of boundary conditions in determining the nature of the flow, we consider Poiseuille flow between two parallel plates located at $y=\pm L$. All types of boundary conditions listed in Table 2 will be considered. We can write boundary conditions (2.49) and (2.51) together as

$$\gamma_*\frac{(|\boldsymbol{v}_{\boldsymbol{\tau}}|-v_*)^+}{|\boldsymbol{v}_{\boldsymbol{\tau}}|}\boldsymbol{v}_{\boldsymbol{\tau}}=\frac{(|\boldsymbol{s}|-s_*)^+}{|\boldsymbol{s}|}\boldsymbol{s} \qquad \text{on } \partial(\mathbb{R}\times(-L,L)) \tag{2.52}$$

requiring that at least one of v_* and s_* is zero. Let us consider a simple shear flow of the Euler/Navier-Stokes fluid (2.36) in domain $\mathbb{R}\times(-L,L)$ for given $L>0$. As shown in Section 2.5 simple shear flows of the Euler/Navier-Stokes fluid are in the form (2.36) or (2.37). Let us

Table 2: Classification of boundary activation of fluid response with the corresponding $|\boldsymbol{v_\tau}|$-$|\boldsymbol{s}|$ diagrams. The last row reflects the usage of the term *slip* in Section 3 to describe a broad class of boundary conditions of the slip type.

free-slip	free-slip/Navier-slip	Navier-slip	no-slip/Navier-slip	no-slip
				
$\boldsymbol{s}=\boldsymbol{0}$	$\lvert\boldsymbol{v_\tau}\rvert \le v_* \iff \boldsymbol{s}=\boldsymbol{0}$	$\boldsymbol{s}\sim\boldsymbol{v_\tau}$	$\lvert\boldsymbol{s}\rvert \le s_* \iff \boldsymbol{v_\tau}=\boldsymbol{0}$	$\boldsymbol{v_\tau}=0$
(2.45)	(2.50) or (2.51)	(2.47)	(2.48) or (2.49)	(2.46)
slip				no-slip

assume symmetry $y_0=0$ in (2.36a); it will be obvious later that the converse is not possible. Normalizing (2.36a) to a given flow rate $Q\in\mathbb{R}$ such that

$$\int_{-L}^{L} u(u)\mathrm{d}y = Q \tag{2.53}$$

we obtain

$$u = -C\left(y^2+\sqrt{2}\delta_*\left|\frac{y}{C}\right|\right)+\frac{Q}{2L}+C\left(\frac{L^2}{3}+\frac{\sqrt{2}\delta_* L}{2|C|}\right), \tag{2.54a}$$

$$p = -2\nu_* Cx + p_0 \tag{2.54b}$$

being defined for any $C\in\mathbb{R}\setminus\{0\}$. It requires a trivial, but tedious, computation to check that simple shear flow of Euler/Navier-Stokes fluid (2.54) solves the balance equations in $\mathbb{R}\times(-L,L)$ together with boundary condition (2.52) on $\{y=\pm L\}$ provided that

$$C=\frac{3Q}{4L^3}\left[\left(1-\frac{\sqrt{2}\delta_* L^2+2v_* L}{|Q|}\right)^+ - \frac{3\nu_*}{3\nu_*+\gamma_* L}\left(1-\frac{\sqrt{2}\delta_* L^2+2v_* L+\frac{2s_* L^2}{3\nu_*}}{|Q|}\right)^+\right] \tag{2.55}$$

and $p_0\in\mathbb{R}$ is arbitrary. If C given by formula (2.55) is zero, then all flows which fulfill (2.37), (2.52), and (2.53) are solutions; more precisely if $C=0$ all the solutions are given by

$$|u'| \le \sqrt{2}\delta_* \quad \text{a.e. in } \mathbb{R}, \tag{2.56a}$$

$$\gamma_*|u| \le \gamma_* v_* \quad \text{a.e. on } \{|y|=L\}, \tag{2.56b}$$

$$\int_{-L}^{L} u\,\mathrm{d}y = Q, \tag{2.56c}$$

$$p = p_0, \tag{2.56d}$$

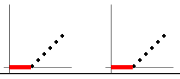
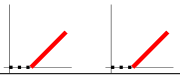
with some Lipschitz continuous $u:[-L,L]\to\mathbb{R}$ and $p_0\in\mathbb{R}$. Family (2.54), (2.55) and family (2.56) represent in fact all possible simple shear flow solutions of motions of the Euler/Navier-Stokes fluid subject to no-slip/Navier-slip or free-slip/Navier-slip boundary conditions (2.52). We summarize combinations of bulk and boundary activation criterions in Table 3.

3 Mathematical analysis of flows of activated Euler fluids

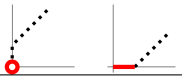
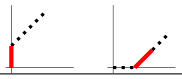
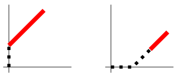
Long-time and large-data existence theory (within the context of weak solutions) for a broad class of fluids described by implicit constitutive relation (2.3) has been developed in [16, 17]. These works deal with internal flows of incompressible fluids with monotone responses, asymptotically behaving as $|\mathbb{S}|=\mathcal{O}\left(|\mathbb{D}|^{r-1}\right)$ as $|\mathbb{D}|\to\infty$ or $|\mathbb{D}|=\mathcal{O}\left(|\mathbb{S}|^{\frac{r}{r-1}-1}\right)$ as $|\mathbb{S}|\to\infty$ with

Table 3: Solutions for simple shear flows of the Euler/Navier-Stokes fluid in combination with different activated and classical boundary conditions. The middle column contains $|\boldsymbol{v}_{\boldsymbol{\tau}}|$-$|\boldsymbol{s}|$ diagrams of boundary response (on the left) and $|\mathbb{D}|$-$|\mathbb{S}|$ diagrams of bulk response (on the right). The solid segments and the circles (colored red in the electronic version) mark the part of the response being attained in the specific case. Note that no-slip/Navier-slip (contrary to free-slip/Navier-slip) admits a mode with the activation threshold exceeded in the bulk with the boundary under activation threshold. Also note that the free-slip condition admits only Euler mode, frictionless solutions. For the Navier-Stokes limit just let $\delta_* := 0$.

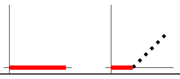
free-slip/Navier-slip: (2.52) & $s_* = 0$

$\lvert Q\rvert \leq \sqrt{2}\delta_* L^2 + 2v_* L$		(2.56)
$\lvert Q\rvert \geq \sqrt{2}\delta_* L^2 + 2v_* L$		(2.54), $C = \frac{\gamma_* L}{3\nu_* + \gamma_* L} \frac{3\left(\lvert Q\rvert - \sqrt{2}\delta_* L^2 - 2v_* L\right)}{4L^3} \frac{Q}{\lvert Q\rvert}$
	bdry bulk	

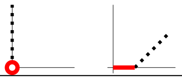
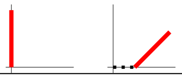
no-slip/Navier-slip: (2.52) & $v_* = 0$

$\lvert Q\rvert \leq \sqrt{2}\delta_* L^2$		(2.56)
$\sqrt{2}\delta_* L^2 \leq \lvert Q\rvert \leq \sqrt{2}\delta_* L^2 + \frac{2s_* L^2}{3\nu_*}$		(2.54), $C = \frac{3\left(\lvert Q\rvert - \sqrt{2}\delta_* L^2\right)}{4L^3} \frac{Q}{\lvert Q\rvert}$
$\lvert Q\rvert \geq \sqrt{2}\delta_* L^2 + \frac{2s_* L^2}{3\nu_*}$		(2.54), $C =$ $\left[\frac{\gamma_* L}{3\nu_* + \gamma_* L} \frac{3\left(\lvert Q\rvert - \sqrt{2}\delta_* L^2\right)}{4L^3} + \frac{3\nu_*}{3\nu_* + \gamma_* L} \frac{s_*}{2\nu_* L}\right] \frac{Q}{\lvert Q\rvert}$
	bdry bulk	

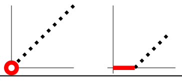
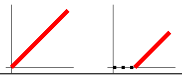
free-slip: (2.52) & $s_* = 0$ & $v_* \to \infty$
free-slip: (2.52) & $s_* = 0$ & $\gamma_* = 0$

$Q \in \mathbb{R}$		(2.56)
	bdry bulk	

no-slip: (2.52) & $v_* = 0$ & $s_* \to \infty$
no-slip: (2.52) & $v_* = 0$ & $\gamma_* \to \infty$

$\lvert Q\rvert \leq \sqrt{2}\delta_* L^2$		(2.56a), $u(\pm L) = 0$, (2.56c), (2.56d)
$\lvert Q\rvert \geq \sqrt{2}\delta_* L^2$		(2.54), $C = \frac{3\left(\lvert Q\rvert - \sqrt{2}\delta_* L^2\right)}{4L^3} \frac{Q}{\lvert Q\rvert}$
	bdry bulk	

Navier-slip: (2.52) & $v_* = 0$ & $s_* = 0$

$\lvert Q\rvert \leq \sqrt{2}\delta_* L^2$		(2.56)
$\lvert Q\rvert \geq \sqrt{2}\delta_* L^2$		(2.54), $C = \frac{\gamma_* L}{3\nu_* + \gamma_* L} \frac{3\left(\lvert Q\rvert - \sqrt{2}\delta_* L^2\right)}{4L^3} \frac{Q}{\lvert Q\rvert}$
	bdry bulk	

$\frac{6}{5} < r < \infty$ (or $\frac{2d}{d+2} < r < \infty$ in general dimension d).[3] In order to get a pressure[4] which is integrable over the space-time cylinder in the unsteady case, the theory is developed with a boundary condition allowing some kind of slip. The overview of the problem concerning the connection between the integrability of the pressure and a specific boundary condition is given in [31]; see also the original studies [43, 44]. Existence theory when the Navier-slip boundary condition is enforced has recently been extended to the stick/slip boundary condition in [22, 21]. The theory for unsteady flows subject to the no-slip boundary condition can be found in a more recent study [13], where the solenoidal Lipschitz approximations of solenoidal Bochner-Sobolev functions are constructed and analyzed (from the point of view of dependence of their mathematical properties on approximation parameters). With such constructions, the analysis of the problem can be performed without introducing the notion of pressure.

In this section we will provide an existence theory for steady and unsteady flows of activated Euler fluids considering various types of behavior after activation and various types of boundary conditions. More specifically, we will study the system[5]

$$\operatorname{div} \boldsymbol{v} = 0 \qquad \text{in } (0,T)\times\Omega, \tag{3.1a}$$

$$\frac{\partial \boldsymbol{v}}{\partial t} + \operatorname{div}(\boldsymbol{v}\otimes\boldsymbol{v}) - \operatorname{div}\mathbb{S} = -\nabla p + \boldsymbol{b} \qquad \text{in } (0,T)\times\Omega, \tag{3.1b}$$

$$\mathbb{S} = 2\nu_* \left(|\mathbb{D}| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}|)\frac{\mathbb{D}}{|\mathbb{D}|} \qquad \text{in } (0,T)\times\Omega, \tag{3.1c}$$

$$\boldsymbol{v}\cdot\boldsymbol{n} = 0 \qquad \text{on } (0,T)\times\partial\Omega, \tag{3.1d}$$

$$\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) = \boldsymbol{0} \qquad \text{on } (0,T)\times\partial\Omega, \tag{3.1e}$$

$$\boldsymbol{v}(0,\cdot) = \boldsymbol{v}_0 \qquad \text{in } \Omega. \tag{3.1f}$$

Here $\mathcal{S} : [0,\infty) \to [0,\infty)$ is supposed to be of the following forms: either

$$\mathcal{S} \equiv 1$$

giving the Euler/Navier-Stokes fluid (2.30), or

$$\mathcal{S}(d) = \left(\tfrac{d}{d_*}\right)^{r-2} \quad \text{or} \quad \mathcal{S}(d) = \Big(A + \left(\tfrac{d}{d_*}\right)^2\Big)^{\frac{r-2}{2}}, \quad A > 0,$$

leading to the Euler/power-law fluid (2.31), or

$$\mathcal{S}(d) = 1 + A\left(\tfrac{d}{d_*}\right)^{r-2}, \quad r > 2, \quad A > 0,$$

leading to the Euler/Ladyzhenskaya fluid.

It is not difficult to verify (see Appendix B) that the graph $\mathcal{G} \subset \mathbb{R}^{3\times3}_{\mathrm{sym}} \times \mathbb{R}^{3\times3}_{\mathrm{sym}}$ defined through

$$(\mathbb{S}, \mathbb{D}) \in \mathcal{G} \text{ if and only if } \mathbb{S} \text{ and } \mathbb{D} \text{ fulfill (3.1c)}$$

is a maximal monotone r-graph, i.e., $\mathcal{G}$ has the following properties:

($\mathcal{G}1$) $(\mathbb{O}, \mathbb{O}) \in \mathcal{G}$;

($\mathcal{G}2$) $(\mathbb{S}_1 - \mathbb{S}_2) : (\mathbb{D}_1 - \mathbb{D}_2) \geq 0$ for all $(\mathbb{S}_1, \mathbb{D}_1) \in \mathcal{G}$ and $(\mathbb{S}_2, \mathbb{D}_2) \in \mathcal{G}$;

($\mathcal{G}3$) if $\mathbb{S}, \mathbb{D} \in \mathbb{R}^{3\times3}_{\mathrm{sym}}$ satisfy $(\mathbb{S} - \tilde{\mathbb{S}}) : (\mathbb{D} - \tilde{\mathbb{D}}) \geq 0$ for all $(\tilde{\mathbb{S}}, \tilde{\mathbb{D}}) \in \mathcal{G}$, then $(\mathbb{S}, \mathbb{D}) \in \mathcal{G}$;

($\mathcal{G}4$) there exist $r \in (1,\infty)$, $\alpha, \beta \in (0,\infty)$ such that $\mathbb{S} : \mathbb{D} \geq \alpha\big(|\mathbb{S}|^{r'} + |\mathbb{D}|^r\big) - \beta$ whenever $(\mathbb{S}, \mathbb{D}) \in \mathcal{G}$ and $\frac{1}{r} + \frac{1}{r'} = 1$.

[3] In fact, in [16, 17] the results are established even in a more general setting replacing the Lebesgue spaces by the Orlicz spaces.

[4] Subtle difference between *thermodynamic pressure*, the *mean normal stress* (the latter usually referred to as pressure in mathematical literature on incompressible fluids), and the Lagrange multiplier is not to be discussed in this paper and we refer interested reader to [59, 77]. Henceforth we refer to the Lagrange multiplier that enforces the incompressibility condition as the "pressure".

[5] We assume that density ρ is constant and we replace $\frac{p}{\rho}$ merely by p throughout the whole section. Note that such p, although customarily called "pressure" in mathematical fluid dynamics literature, is actually the mean normal stress scaled by the (constant) density.

Suitable choices of the function $\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v_\tau})$ cover the boundary conditions (2.45), (2.46), (2.47), (2.49), (2.51). Analogous to the above setting, we require that the graph $\mathcal{B} \subset \mathbb{R}^3 \times \mathbb{R}^3$ defined through

$$(\boldsymbol{s}, \boldsymbol{v}) \in \mathcal{B} \Leftrightarrow \boldsymbol{s} \text{ and } \boldsymbol{v} \text{ fulfill (3.1e)}$$

is a maximal monotone 2-graph, i.e., $\mathcal{B}$ has the following properties:

($\mathcal{B}$1) $(\boldsymbol{0}, \boldsymbol{0}) \in \mathcal{B}$;

($\mathcal{B}$2) $(\boldsymbol{s}_1 - \boldsymbol{s}_2) \cdot (\boldsymbol{v}_1 - \boldsymbol{v}_2) \geq 0$ for all $(\boldsymbol{s}_1, \boldsymbol{v}_1) \in \mathcal{B}$ and $(\boldsymbol{s}_2, \boldsymbol{v}_2) \in \mathcal{B}$;

($\mathcal{B}$3) if $\boldsymbol{s}, \boldsymbol{v} \in \mathbb{R}^3$ satisfy $(\boldsymbol{s} - \tilde{\boldsymbol{s}}) \cdot (\boldsymbol{v} - \tilde{\boldsymbol{v}}) \geq 0$ for all $(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{v}}) \in \mathcal{B}$, then $(\boldsymbol{s}, \boldsymbol{v}) \in \mathcal{B}$;

($\mathcal{B}$4) there are $\tilde{\alpha}, \tilde{\beta} \in (0, \infty)$ such that $\boldsymbol{s} \cdot \boldsymbol{v} \geq \tilde{\alpha} \left(|\boldsymbol{s}|^2 + |\boldsymbol{v}|^2\right) - \tilde{\beta}$ for all $(\boldsymbol{s}, \boldsymbol{v}) \in \mathcal{B}$.

The requirement ($\mathcal{B}$4) can be easily verified for the boundary conditions (2.47), (2.49) and (2.51).

Note that there is no boundary term in the weak formulation of the problem in the case of the free-slip condition (2.45); this condition does not invalidate the analysis. On the other hand the no-slip boundary condition (2.46) needs to be treated separately.

Apart from the general purpose of this paper we are further motivated to study the problem (3.1) for the following reasons.

1. The most studied systems of PDEs (partial differential equations) in fluid mechanics are the Euler equations (when $\mathbb{S} = \mathbb{O}$, or $\delta_* \to \infty$ in (3.1c)) and the Navier-Stokes equations (when $\mathbb{S} = 2\nu_* \mathbb{D}$, or $\delta_* = 0$ and $\mathcal{S} \equiv 1$ in (3.1c)). The system of PDEs considered here is placed between them, as $\delta_* \in (0, \infty)$. While (3.1a)–(3.1c) can, particularly for δ_* large, share several features associated with the physics of the Euler fluid (or the Euler equations), we will document that the mathematical properties of the flows described by (3.1) are similar to those described by the Navier-Stokes equations. This is important as recent achievements in the mathematical theory of the Euler equations considered in a reasonable physical setting show that the equations exhibit pathological solutions within the framework of weak solutions with bounded (kinetic) energy (see [24, 82]).

   Fluids described by (3.1c) seem to have been completely overlooked both in physics and mathematical fluid dynamics literature; this may well be due to the fact that such behavior has not been observed. Below, we will focus on filling this lacuna and on developing the mathematical foundations associated with the problem (3.1).

2. It is worth noticing that the activated Euler fluids characterized by (3.1c) represent the models dual to the Bingham fluids that are obtained by interchanging the role of $\mathbb{D}$ and $\mathbb{S}$ in (3.1c). A mathematical theory for Bingham fluids, in the spirit of the theory developed here, is given in [21, 22, 63], where the reader can also find more references concerning the earlier results on the analysis of flows of the Bingham fluids and their generalizations.

3. The set-up of the problem considered here will be also used to show how different types of boundary conditions can be treated (while restricting ourselves to internal flows). We will also focus on the relation between the considered boundary conditions and the properties of the mean normal stress p.

4. Since the operator $-\operatorname{div} \mathbb{S}$ is elliptic and degenerates for $|\mathbb{S}| \leq \delta_*$, the theory presented below can be viewed as an approach for studying degenerate problems.

5. Finally, the constitutive relation (3.1c) is regularized by

$$\mathbb{S}^\epsilon(\mathbb{D}) = \left(\epsilon |\mathbb{D}|^{q-2} + 2\nu_* \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|} \mathcal{S}(|\mathbb{D}|)\right) \mathbb{D}$$

   with $\epsilon > 0$ and $q \geq 2$ large enough. This explicit regularization allows us to proceed explicitly in the subsequent analysis.

3.1 Function spaces

In what follows we assume that $\Omega \subset \mathbb{R}^3$ is a domain, i.e., a bounded open connected set. For $1 \leq p \leq \infty$, $k \in \mathbb{N}$, $L^p(\Omega)$ and $W^{k,p}(\Omega)$ denote the standard Lebesgue and Sobolev space respectively, i.e., spaces of measurable functions of finite norm

$$\begin{aligned} \|f\|_{L^p(\Omega)} &= \|f\|_{p,\Omega} &&= \Big(\int_\Omega |f|^p\Big)^{\frac{1}{p}}, \\ \|f\|_{W^{k,p}(\Omega)} &= \|f\|_{k,p,\Omega} &&= \sum_{j=0}^{k} \||\nabla^j f|\|_{p,\Omega} \quad |\nabla^j f| = \Big(\sum_{|\alpha|=j} |\mathrm{D}^\alpha f|^2\Big)^{\frac{1}{2}} \end{aligned}$$

respectively. When there is no risk of confusion the subscript Ω can be omitted. Often we will use symbol $L^p(\Omega)^{3\times 3}_{\mathrm{sym}}$ to denote functions with values in symmetric tensors with L^p-integrable components. Bold-face symbols $\mathbf{W}^{k,p}$ and $\mathbf{L}^p$ will denote vector-valued Sobolev and Lebesgue functions respectively. Parentheses $(\cdot,\cdot)_\Omega$ will denote duality pairing in $L^p(\Omega)$ and $L^{\frac{p}{p-1}}(\Omega)$ including vector and tensor-valued case; the subscript Ω will be typically dropped where is no danger of confusion. Analogously, angle brackets $\langle\cdot,\cdot\rangle_{V^*,V}$ denote a duality paring between spaces V^* and V, where V^* denotes the dual of V; the subscript can be omitted. For any vector-valued function $\boldsymbol{v}$, the symmetric part of the gradient is defined through $\mathbb{D}\boldsymbol{v} := \frac{1}{2}\big(\nabla\boldsymbol{v} + (\nabla\boldsymbol{v})^\top\big)$. We use notation $L^q(0,T;X)$ and $\mathcal{C}^k(I;X)$ to denote Bochner spaces of functions with values in the Banach space X and k-times continuously differentiable functions on interval $I \subset \mathbb{R}$ with values in X respectively; $\mathcal{C}^k_0(I;X)$ denotes functions from $\mathcal{C}^k(I;X)$ which are compactly supported in I.

The function spaces relevant to the problems that are being investigated vary depending on the type of boundary conditions. Two cases are being considered separately. First, the case of the no-slip boundary condition (the no-slip case, in short) and then other boundary conditions that involve various kinds of slipping mechanisms (the slip case, in short), cf. Table 2.

3.1.1 No-slip case

We consider the space of compactly supported smooth functions and its subspace of solenoidal functions:

$$\boldsymbol{\mathcal{C}}^\infty_0 := \big\{\boldsymbol{v} : \Omega \to \mathbb{R}^3;\ \boldsymbol{v} \text{ smooth};\ \operatorname{supp}\boldsymbol{v} \subset \Omega\big\}, \quad \boldsymbol{\mathcal{C}}^\infty_{0,\mathrm{div}} := \{\boldsymbol{v} \in \boldsymbol{\mathcal{C}}^\infty_0;\ \operatorname{div}\boldsymbol{v} = 0\}$$

and their closures in L^p-norm, $W^{1,p}$-norm (with $1 < p < \infty$) and $W^{3,2}$-norm:

$$\begin{aligned} \mathbf{W}^{1,p}_0 &:= \overline{\boldsymbol{\mathcal{C}}^\infty_0}^{\|\cdot\|_{W^{1,p}}}, \\ \mathbf{L}^p_{\boldsymbol{n},\mathrm{div}} &:= \overline{\boldsymbol{\mathcal{C}}^\infty_{0,\mathrm{div}}}^{\|\cdot\|_{L^p}}, \\ \mathbf{W}^{1,p}_{0,\mathrm{div}} &:= \overline{\boldsymbol{\mathcal{C}}^\infty_{0,\mathrm{div}}}^{\|\cdot\|_{W^{1,p}}}, \end{aligned} \qquad \begin{aligned} \mathbf{W}^{3,2}_0 &:= \overline{\boldsymbol{\mathcal{C}}^\infty_0}^{\|\cdot\|_{W^{3,2}}}, \\ \mathbf{W}^{3,2}_{0,\mathrm{div}} &:= \overline{\boldsymbol{\mathcal{C}}^\infty_{0,\mathrm{div}}}^{\|\cdot\|_{W^{3,2}}}. \end{aligned}$$

As a consequence of the Poincaré and Korn inequalities, see [2, Corollary 6.31], [26, Theorem 5.15], the following norms are equivalent on $\mathbf{W}^{1,p}_0$ (and $\mathbf{W}^{1,p}_{0,\mathrm{div}} \subset \mathbf{W}^{1,p}_0$) for $1 < p < \infty$:

$$\|\mathbb{D}\boldsymbol{v}\|_p \leq \|\nabla\boldsymbol{v}\|_p \leq \|\boldsymbol{v}\|_{1,p} \leq C_{\mathrm{P}}\|\nabla\boldsymbol{v}\|_p \leq C_{\mathrm{P}}C_{\mathrm{K}}\|\mathbb{D}\boldsymbol{v}\|_p \quad \text{for all } \boldsymbol{v} \in \mathbf{W}^{1,p}_0, \tag{3.2}$$

with $\|\nabla\boldsymbol{v}\|_p := \||\nabla\boldsymbol{v}|\|_p$, $\|\mathbb{D}\boldsymbol{v}\|_p := \||\mathbb{D}\boldsymbol{v}|\|_p$; the constant $C_{\mathrm{P}} > 0$ that appears due to the Poincaré inequality depends on p and Ω, while the constant $C_{\mathrm{K}} > 0$ that appears due to the Korn inequality depends only on p.

Note that for a domain Ω (without further regularity assumption on the smoothness of $\partial\Omega$) we have the embedding $\mathbf{W}^{3,2}_{0,\mathrm{div}} \hookrightarrow \mathbf{W}^{3,2}_0 \hookrightarrow W^{1,\infty}(\Omega)^3$. If additionally Ω is a $C^{0,1}$ domain, i.e., Ω is a domain with Lipschitz boundary $\partial\Omega$, then the following characterization holds true:

$$\begin{aligned} \mathbf{W}^{1,p}_0 &= \big\{\boldsymbol{v} \in W^{1,p}(\Omega)^3;\ \boldsymbol{v} = \mathbf{0} \text{ on } \partial\Omega \text{ in the sense of traces}\big\}, \\ \mathbf{W}^{1,p}_{0,\mathrm{div}} &= \Big\{\boldsymbol{v} \in \mathbf{W}^{1,p}_0;\ \operatorname{div}\boldsymbol{v} = 0\Big\}; \end{aligned} \tag{3.3}$$

moreover we use (3.3) as a definition of $\mathbf{W}^{1,\infty}_0$ and $\mathbf{W}^{1,\infty}_{0,\mathrm{div}}$ in the case that $p = \infty$ and Ω is a $C^{0,1}$ domain.

We can occasionally denote the norm on $\big(\mathbf{W}^{1,p}_0\big)^*$, the topological dual of $\mathbf{W}^{1,p}_0$, by $\|\cdot\|_{-1,p'}$.

3.1.2 Slip case

Here we assume Ω is a $C^{0,1}$ domain. We denote by $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^3$ the unit outer normal vector to $\partial\Omega$. The space of smooth vector-valued functions with vanishing normal component on the boundary and its solenoidal subspace are then introduced through:

$$\begin{aligned} \boldsymbol{\mathcal{C}}^\infty_{\boldsymbol{n}} &:= \left\{\boldsymbol{v} : \Omega \to \mathbb{R}^3;\ \boldsymbol{v} \text{ smooth};\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\right\}, \\ \boldsymbol{\mathcal{C}}^\infty_{\boldsymbol{n},\mathrm{div}} &:= \left\{\boldsymbol{v} \in \boldsymbol{\mathcal{C}}^\infty_{\boldsymbol{n}};\ \operatorname{div}\boldsymbol{v} = 0\right\}. \end{aligned}$$

Since $\partial\Omega$ is Lipschitz we can define the following spaces with vanishing normal trace:

$$\begin{aligned} \mathbf{W}^{3,2}_{\boldsymbol{n}} &:= \left\{\boldsymbol{v} \in W^{3,2}(\Omega)^3;\ \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\right\}, \\ \mathbf{W}^{3,2}_{\boldsymbol{n},\mathrm{div}} &:= \mathbf{W}^{3,2}_{\boldsymbol{n}} \cap \mathbf{L}^2_{\boldsymbol{n},\mathrm{div}}, \end{aligned}$$

and subsequently, for $1 < p \le \infty$,

$$\mathbf{W}^{1,p}_{\boldsymbol{n}} := \overline{\mathbf{W}^{3,2}_{\boldsymbol{n}}}^{\|\cdot\|_{W^{1,p}}}, \quad \mathbf{W}^{1,p}_{\boldsymbol{n},\mathrm{div}} := \overline{\mathbf{W}^{3,2}_{\boldsymbol{n},\mathrm{div}}}^{\|\cdot\|_{W^{1,p}}}.$$

The condition $\boldsymbol{v}\cdot\boldsymbol{n} = 0$ on $\partial\Omega$ for Ω bounded is sufficient for validity of the Poincaré inequality:[6] for $1 < p < \infty$ there exists $C_{\mathrm{P}} > 0$ depending on p and Ω such that

$$\|\nabla\boldsymbol{v}\|_p \le \|\boldsymbol{v}\|_{1,p} \le C_{\mathrm{P}}\|\nabla\boldsymbol{v}\|_p \qquad \text{for all } \boldsymbol{v} \in \mathbf{W}^{1,p}_{\boldsymbol{n}}. \tag{3.4}$$

For the steady problem we will consider two inequalities of the Korn type depending on whether the type of considered boundary conditions leads to the control of the trace of $\boldsymbol{v}$ on the boundary or not.[7] In the first case, it follows from [19, Lemma 1.11] and (3.4): for $1 < p < \infty$ there exists $C_{\mathrm{K}} > 0$ depending on p and Ω such that

$$\|\nabla\boldsymbol{v}\|_p \le C_{\mathrm{K}}\big(\|\mathbb{D}\boldsymbol{v}\|_p + \|\boldsymbol{v}\|_{2,\partial\Omega}\big) \qquad \text{for all } \boldsymbol{v} \in \mathbf{W}^{1,p}_{\boldsymbol{n}} \text{ with } \boldsymbol{v}_{\boldsymbol{\tau}} \in \mathbf{L}^2(\partial\Omega). \tag{3.5}$$

The second situation when $\boldsymbol{s} = \mathbf{0}$ on $\partial\Omega$ requires us to rule out domains that admit nontrivial rigid motions. We say that Ω is axisymmetric if there exists a rigid body motion tangential to boundary, i.e., there is $\boldsymbol{v} \in \mathbf{W}^{1,\infty}_{\boldsymbol{n}}$ with $\mathbb{D}\boldsymbol{v} = \mathbb{O}$ and $\nabla\boldsymbol{v} \ne \mathbb{O}$ constant in Ω. In the other words, there is $\boldsymbol{v} \in \mathbf{W}^{1,\infty}_{\boldsymbol{n}}$ of the form $\boldsymbol{v}(\boldsymbol{x}) = \mathbb{Q}(\boldsymbol{x} - \boldsymbol{x}_0)$ for some $\mathbb{Q} \subset \mathbb{R}^{3\times 3}$ non-zero skew-symmetric matrix and constant $\boldsymbol{x}_0 \in \mathbb{R}^3$. From [9, Theorem 11, Remark 12] it follows that if Ω is not axisymmetric and $1 < p < \infty$ there exists $C_{\mathrm{K}} > 0$ depending on Ω and p such that

$$\|\nabla\boldsymbol{v}\|_p \le C_{\mathrm{K}}\|\mathbb{D}\boldsymbol{v}\|_p \qquad \text{for all } \boldsymbol{v} \in \mathbf{W}^{1,p}_{\boldsymbol{n}}. \tag{3.6}$$

For the unsteady case we will use the following Korn-type inequality:[8]

$$\|\nabla\boldsymbol{v}\|_p \le C_{\mathrm{K}}\big(\|\mathbb{D}\boldsymbol{v}\|_p + \|\boldsymbol{v}\|_1\big) \qquad \text{for all } \boldsymbol{v} \in \mathbf{W}^{1,p}_{\boldsymbol{n}}. \tag{3.8}$$

[6] To verify it, assume for the sake of contradicion, that there is $\{\boldsymbol{v}^j\}_{j=1}^\infty \subset \mathbf{W}^{1,p}_{\boldsymbol{n}}$ with $\|\nabla\boldsymbol{v}^j\|_p \to 0$ and $\|\boldsymbol{v}^j\|_{1,p} = 1$. Relying on the compact Sobolev embedding, it follows that there is a (not relabeled) subsequence $\{\boldsymbol{v}^j\}_{j=1}^\infty$ which converges strongly in $L^p(\Omega)^3$ to some $\boldsymbol{v} \in L^p(\Omega)^3$. This implies that $\{\boldsymbol{v}^j\}_{j=1}^\infty$ is a Cauchy sequence in $\mathbf{W}^{1,p}_{\boldsymbol{n}}$ and converges in $\mathbf{W}^{1,p}_{\boldsymbol{n}}$ to $\boldsymbol{v} \in \mathbf{W}^{1,p}_{\boldsymbol{n}}$ with $\nabla\boldsymbol{v} = \mathbb{O}$. Hence $\boldsymbol{v}$ is constant and by virtue of the boundedness of Ω and the boundary condition it follows that $\boldsymbol{v} = \mathbf{0}$, which is a contradiction.

[7] A priori estimates for a steady problem subject to a slip-type condition given by a maximal monotone 2-graph with ($\mathcal{B}4$) will ensure control of $\|\boldsymbol{v}_{\boldsymbol{\tau}}\|_{2,\partial\Omega}$. On the other hand, under the free-slip condition (2.45) there is no a priori control over the tangential velocity $\|\boldsymbol{v}_{\boldsymbol{\tau}}\|_{2,\partial\Omega}$ and rigid motions, if admissible in $\mathbf{W}^{1,\infty}_{\boldsymbol{n}}$, prevent one to obtain a steady solution.

[8] This is a consequence of another Korn-type inequality:

$$\|\nabla\boldsymbol{v}\|_p \le C'_{\mathrm{K}}\big(\|\mathbb{D}\boldsymbol{v}\|_p + \|\boldsymbol{v}\|_p\big) \qquad \text{for all } \boldsymbol{v} \in \mathbf{W}^{1,p}(\Omega); \tag{3.7}$$

see [56, Theorem 1.10]. To verify (3.8), assume that there is $\{\boldsymbol{v}^j\}_{j=1}^\infty \subset \mathbf{W}^{1,p}_{\boldsymbol{n}}$ such that $\|\nabla\boldsymbol{v}^j\|_p = 1$ and $\|\mathbb{D}\boldsymbol{v}^j\|_p + \|\boldsymbol{v}^j\|_1 \to 0$. Poincaré inequality (3.4) implies that $\{\boldsymbol{v}^j\}_{j=1}^\infty$ is bounded in $\mathbf{W}^{1,p}_{\boldsymbol{n}}$. By virtue of its reflexivity there is a (not relabeled) subsequence such that $\boldsymbol{v}^j \rightharpoonup \boldsymbol{v}$ weakly in $\mathbf{W}^{1,p}_{\boldsymbol{n}}$ and by the Sobolev embedding $\boldsymbol{v}^j \to \boldsymbol{v}$ in $\mathbf{L}^p$. On the other hand $\boldsymbol{v}^j \to \mathbf{0}$ in $\mathbf{L}^1$ hence by uniqueness of the limit we conclude $\boldsymbol{v} = \mathbf{0}$. Summarizing, the right-hand side of (3.7) (with $\boldsymbol{v}^j$ in place of $\boldsymbol{v}$) goes to zero but the left-hand side is equal to unity, which is a contradiction.

We can see that (3.8) in fact holds independently of the considered boundary condition, i.e., for all $\mathbf{W}^{1,p}$.

3.2 Analysis of steady flows

In this section, we investigate internal flows that are independent of time. Under such circumstances, the governing system of equations takes the form

$$\begin{aligned}
\operatorname{div}\boldsymbol{v} &= 0 && \text{in } \Omega, && \text{(3.9a)}\\
\operatorname{div}(\boldsymbol{v}\otimes\boldsymbol{v}-\mathbb{S}) &= -\nabla p+\boldsymbol{b} && \text{in } \Omega, && \text{(3.9b)}\\
(\mathbb{S},\mathbb{D}\boldsymbol{v}) &\in \mathcal{G} && \text{in } \Omega, && \text{(3.9c)}\\
\boldsymbol{v}\cdot\boldsymbol{n} &= 0 && \text{on } \partial\Omega, && \text{(3.9d)}
\end{aligned}$$

where $\mathcal{G}$ is a maximal monotone r-graph, fulfilling ($\mathcal{G}1$)–($\mathcal{G}4$), of the form

$$\mathcal{G} := \left\{(\mathbb{S},\mathbb{D}) \in \mathbb{R}^{3\times3}_{\mathrm{sym}} \times \mathbb{R}^{3\times3}_{\mathrm{sym}};\ \mathbb{S} = 2\nu_*\left(|\mathbb{D}|-\delta_*\right)^+ \mathcal{S}(|\mathbb{D}|)\tfrac{\mathbb{D}}{|\mathbb{D}|}\right\}. \tag{3.10}$$

We will distinguish two cases depending on the class of boundary conditions considered. The first case concerns the no-slip condition, i.e.,

$$\boldsymbol{v}_{\boldsymbol{\tau}} = \mathbf{0} \qquad \text{on } \partial\Omega. \tag{3.11}$$

The second case includes all other boundary conditions involving tangential part of the normal traction; it refers to either

$$\boldsymbol{s} = \mathbf{0} \qquad \text{on } \partial\Omega$$

or

$$(\boldsymbol{s},\boldsymbol{v}_{\boldsymbol{\tau}}) \in \mathcal{B} \qquad \text{on } \partial\Omega,$$

where $\mathcal{B}$ is a maximal monotone 2-graph fulfilling ($\mathcal{B}1$)–($\mathcal{B}4$).

3.2.1 No-slip case

Let $\Omega \subset \mathbb{R}^3$ be a domain, $r \geq \frac{6}{5}$, $\boldsymbol{b} \in \left(\mathbf{W}^{1,r}_0\right)^*$ and $\mathcal{G}$ be a maximal monotone r-graph specified in (3.10). We say that

$$(\boldsymbol{v},\mathbb{S}) \in \mathbf{W}^{1,r}_{0,\mathrm{div}} \times L^{r'}(\Omega)^{3\times3}_{\mathrm{sym}}$$

is a weak solution to (3.9), (3.10), (3.11) if

$$(\mathbb{S},\mathbb{D}\boldsymbol{\varphi}) - (\boldsymbol{v}\otimes\boldsymbol{v},\nabla\boldsymbol{\varphi}) = \langle\boldsymbol{b},\boldsymbol{\varphi}\rangle \qquad \text{for all } \boldsymbol{\varphi} \in \boldsymbol{\mathcal{C}}^{\infty}_{0,\mathrm{div}} \tag{3.12}$$

and

$$(\mathbb{S},\mathbb{D}\boldsymbol{v}) \in \mathcal{G} \qquad \text{a.e. in } \Omega; \tag{3.13}$$

equivalently, we can require that (3.12) holds for all $\boldsymbol{\varphi} \in \mathbf{W}^{1,\frac{3r}{5r-6}}_{0,\mathrm{div}} \cap \mathbf{W}^{1,r}_{0,\mathrm{div}}$.

Theorem 3.1. *Let $\Omega \subset \mathbb{R}^3$ be a domain. Let $r > \frac{6}{5}$, $\boldsymbol{b} \in \left(\mathbf{W}^{1,r}_0\right)^*$ and $\mathcal{G}$ be a maximal monotone r-graph of the form* (3.10). *Then there is a weak solution $(\boldsymbol{v},\mathbb{S}) \in \mathbf{W}^{1,r}_{0,\mathrm{div}} \times L^{r'}(\Omega)^{3\times3}_{\mathrm{sym}}$ to* (3.9), (3.11) *which fulfills* (3.13) *and*

$$\begin{aligned}
(\mathbb{S},\mathbb{D}\boldsymbol{\varphi}) - (\boldsymbol{v}\otimes\boldsymbol{v},\mathbb{D}\boldsymbol{\varphi}) &= \langle\boldsymbol{b},\boldsymbol{\varphi}\rangle\\
&\textit{for all } \boldsymbol{\varphi} \in \begin{cases} \mathbf{W}^{1,r}_{0,\mathrm{div}} & \textit{if } r \geq \frac{9}{5},\\ \mathbf{W}^{1,\frac{3r}{5r-6}}_{0,\mathrm{div}} & \textit{if } r \in \left(\frac{6}{5},\frac{9}{5}\right). \end{cases}
\end{aligned} \tag{3.14}$$

In addition, if Ω is a $C^{0,1}$ domain then there is

$$p \in \begin{cases} L^{r'}(\Omega) & \textit{if } r \geq \frac{9}{5}\\ L^{\frac{3r}{2(3-r)}}(\Omega) & \textit{if } r \in \left(\frac{6}{5},\frac{9}{5}\right) \end{cases}, \qquad \int_\Omega p\,\mathrm{d}x = 0, \tag{3.15}$$

such that

$$\begin{aligned}
(\mathbb{S},\mathbb{D}\boldsymbol{\varphi}) - (\boldsymbol{v}\otimes\boldsymbol{v},\nabla\boldsymbol{\varphi}) &= (p,\operatorname{div}\boldsymbol{\varphi}) + \langle\boldsymbol{b},\boldsymbol{\varphi}\rangle\\
&\textit{for all } \boldsymbol{\varphi} \in \begin{cases} \mathbf{W}^{1,r}_{0} & \textit{if } r \geq \frac{9}{5},\\ \mathbf{W}^{1,\frac{3r}{5r-6}}_{0} & \textit{if } r \in \left(\frac{6}{5},\frac{9}{5}\right). \end{cases}
\end{aligned} \tag{3.16}$$

Proof. **The case $r \geq \frac{9}{5}$.** Since $\mathbf{W}^{1,r}_{0,\mathrm{div}}$ is separable, there is a countable basis denoted by $\{\boldsymbol{\omega}^r\}_{r=1}^{\infty}$. For $N \in \mathbb{N}$ arbitrary, but fixed, we first look for the vector $\boldsymbol{c}^N = \left(c_1^N, \dots, c_N^N\right) \in \mathbb{R}^N$ such that

$$\boldsymbol{v}^N(x) := \sum_{r=1}^{N} c_r^N \boldsymbol{\omega}^r(x)$$

satisfies the system of N nonlinear equations

$$\left(\mathbb{S}^N, \mathbb{D}\boldsymbol{\omega}^r\right) - \left(\boldsymbol{v}^N \otimes \boldsymbol{v}^N, \nabla \boldsymbol{\omega}^r\right) = \langle \boldsymbol{b}, \boldsymbol{\omega}^r \rangle, \quad r = 1, \dots, N, \tag{3.17}$$

where

$$\mathbb{S}^N := \mathbb{S}\left(\mathbb{D}\boldsymbol{v}^N\right) := 2\nu_* \left(\left|\mathbb{D}\boldsymbol{v}^N\right| - \delta_*\right)^+ \mathcal{S}\left(\left|\mathbb{D}\boldsymbol{v}^N\right|\right) \frac{\mathbb{D}\boldsymbol{v}^N}{|\mathbb{D}\boldsymbol{v}^N|}. \tag{3.18}$$

Introducing the (continuous) mapping $\boldsymbol{P}^N : \mathbb{R}^N \to \mathbb{R}^N$ as

$$\Big(\boldsymbol{P}^N\big(\boldsymbol{c}^N\big)\Big)_r := \left(\mathbb{S}^N, \mathbb{D}\boldsymbol{\omega}^r\right) - \left(\boldsymbol{v}^N \otimes \boldsymbol{v}^N, \nabla \boldsymbol{\omega}^r\right) - \langle \boldsymbol{b}, \boldsymbol{\omega}^r \rangle, \quad r = 1, \dots, N,$$

then

$$\boldsymbol{P}^N\big(\boldsymbol{c}^N\big) \cdot \boldsymbol{c}^N = \left(\mathbb{S}^N, \mathbb{D}\boldsymbol{v}^N\right) - \left\langle \boldsymbol{b}, \boldsymbol{v}^N \right\rangle. \tag{3.19}$$

It follows from ($\mathcal{G}4$) and (3.18) that

$$\boldsymbol{P}^N\big(\boldsymbol{c}^N\big) \cdot \boldsymbol{c}^N > 0 \quad \text{for } |\boldsymbol{c}^N| \text{ sufficiently large.} \tag{3.20}$$

As a consequence of Brouwer's fixed-point theorem (see [51, p. 53]), (3.20) implies the existence of $\boldsymbol{c}^N$ fulfilling $\boldsymbol{P}^N\big(\boldsymbol{c}^N\big) = \boldsymbol{0}$, i.e., (3.17) holds, and, by (3.19),

$$\left(\mathbb{S}^N, \mathbb{D}\boldsymbol{v}^N\right) = \left\langle \boldsymbol{b}, \boldsymbol{v}^N \right\rangle. \tag{3.21}$$

This together with ($\mathcal{G}4$), (3.2) and Young's inequality leads to

$$\|\mathbb{S}^N\|_{r'} + \|\nabla \boldsymbol{v}^N\|_r \leq c_1 \|\boldsymbol{b}\|_{\left(\mathbf{W}^{1,r}_0\right)^*} + c_2.$$

This implies the existence of $\boldsymbol{v} \in \mathbf{W}^{1,r}_{0,\mathrm{div}}$ and $\mathbb{S} \in L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}$ such that for suitable (not relabeled) subsequences

$$\boldsymbol{v}^N \rightharpoonup \boldsymbol{v} \qquad \text{weakly in } \mathbf{W}^{1,r}_{0,\mathrm{div}}, \tag{3.22a}$$

$$\mathbb{D}\boldsymbol{v}^N \rightharpoonup \mathbb{D}\boldsymbol{v} \qquad \text{weakly in } L^r(\Omega)^{3\times 3}_{\mathrm{sym}}, \tag{3.22b}$$

$$\mathbb{S}^N \rightharpoonup \mathbb{S} \qquad \text{weakly in } L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}, \tag{3.22c}$$

as $N \to \infty$. Consequently, as $W^{1,q}_0(\Omega)$ is compactly embedded into $L^2(\Omega)$ for any $q > \frac{6}{5}$, we also have

$$\boldsymbol{v}^N \to \boldsymbol{v} \quad \text{strongly in } L^2(\Omega)^3 \text{ as } N \to \infty.$$

Then, (3.17) leads to, for $r \geq \frac{9}{5}$,

$$(\mathbb{S}, \mathbb{D}\boldsymbol{\omega}^s) - (\boldsymbol{v} \otimes \boldsymbol{v}, \nabla \boldsymbol{\omega}^s) = \langle \boldsymbol{b}, \boldsymbol{\omega}^s \rangle, \quad s = 1, 2, \dots. \tag{3.23}$$

Note that the restriction $r \geq \frac{9}{5}$ is due to the requirement that for $s \in \mathbb{N}$ arbitrary

$$\int_\Omega (\boldsymbol{v} \otimes \boldsymbol{v}) : \nabla \boldsymbol{\omega}^s \, \mathrm{d}x < \infty \quad \text{for } \boldsymbol{v}, \boldsymbol{\omega}^s \in \mathbf{W}^{1,r}_0.$$

Hence (3.23) implies that

$$(\mathbb{S}, \mathbb{D}\boldsymbol{\omega}) - (\boldsymbol{v} \otimes \boldsymbol{v}, \nabla \boldsymbol{\omega}) = \langle \boldsymbol{b}, \boldsymbol{\omega} \rangle \qquad \text{for all } \boldsymbol{\omega} \in \mathbf{W}^{1,r}_{0,\mathrm{div}}, \tag{3.24}$$

which completes the proof of (3.14) for $r \geq \frac{9}{5}$. Taking $\boldsymbol{\omega} = \boldsymbol{v}$ in (3.24) one obtains

$$(\mathbb{S}, \mathbb{D}\boldsymbol{v}) = \langle \boldsymbol{b}, \boldsymbol{v} \rangle . \tag{3.25}$$

Taking the limit with $N \to \infty$ in (3.21), we conclude from (3.25) and (3.22a) that

$$\lim_{N\to\infty} \left(\mathbb{S}^N, \mathbb{D}\boldsymbol{v}^N\right) = (\mathbb{S}, \mathbb{D}\boldsymbol{v}) .$$

In virtue of the graph convergence lemma (see Lemma A.6 in Appendix) this implies together with (3.22) that $(\mathbb{S}, \mathbb{D}\boldsymbol{v}) \in \mathcal{G}$, i.e., (3.13) holds.

The case $r \in \left(\frac{6}{5}, \frac{9}{5}\right)$. In this case, we consider, for $\epsilon > 0$, the following approximating problem:

$$\begin{aligned} -\operatorname{div}(\mathbb{S} + \epsilon \mathbb{D}\boldsymbol{v} - \boldsymbol{v} \otimes \boldsymbol{v}) &= -\nabla p + \boldsymbol{b} && \text{in } \Omega, \\ \operatorname{div} \boldsymbol{v} &= 0 && \text{in } \Omega, \\ (\mathbb{S}, \mathbb{D}\boldsymbol{v}) &\in \mathcal{G} && \text{in } \Omega, \\ \boldsymbol{v} &= \boldsymbol{0} && \text{on } \partial\Omega. \end{aligned} \tag{3.26}$$

For fixed ϵ, the existence of a weak solution to (3.26) follows from the above proof for the case $r \geq \frac{9}{5}$. More precisely, following step-by-step the proof of existence via the Galerkin method used above we can show that, for $\epsilon > 0$ fixed, there is $(\boldsymbol{v}^\epsilon, \mathbb{S}^\epsilon) \in \mathbf{W}^{1,2}_{0,\mathrm{div}} \times L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}$ such that[9]

$$(\mathbb{S}^\epsilon + \epsilon \mathbb{D}\boldsymbol{v}^\epsilon - \boldsymbol{v}^\epsilon \otimes \boldsymbol{v}^\epsilon, \mathbb{D}\boldsymbol{\varphi}) = \langle \boldsymbol{b}, \boldsymbol{\varphi} \rangle \text{ for all } \boldsymbol{\varphi} \in \mathbf{W}^{1,2}_{0,\mathrm{div}}, \tag{3.27a}$$

$$(\mathbb{S}^\epsilon, \mathbb{D}\boldsymbol{v}^\epsilon) \in \mathcal{G} \text{ a.e. in } \Omega. \tag{3.27b}$$

Moreover, taking $\boldsymbol{\varphi} = \boldsymbol{v}^\epsilon$ in (3.27a),

$$(\mathbb{S}^\epsilon, \mathbb{D}\boldsymbol{v}^\epsilon) + \epsilon \left\|\mathbb{D}\boldsymbol{v}^\epsilon\right\|_2^2 = \langle \boldsymbol{b}, \boldsymbol{v}^\epsilon \rangle .$$

The last identity together with the assumption ($\mathcal{G}4$) implies the following estimate

$$\|\mathbb{S}^\epsilon\|_{r'}^{r'} + \|\mathbb{D}\boldsymbol{v}^\epsilon\|_r^r + \epsilon \|\mathbb{D}\boldsymbol{v}^\epsilon\|_2^2 \leq \|\boldsymbol{b}\|^{r'}_{\left(\mathbf{W}^{1,r}_0\right)^*} + C. \tag{3.28}$$

This implies the existence of $(\boldsymbol{v}, \mathbb{S}) \in \mathbf{W}^{1,r}_{0,\mathrm{div}} \times L^{r'}(\Omega)^{3\times3}_{\mathrm{sym}}$ such that, for a suitable vanishing subsequence $\{\epsilon_n\}_{n=1}^\infty$ and $(\boldsymbol{v}^n, \mathbb{S}^n) := (\boldsymbol{v}^{\epsilon_n}, \mathbb{S}^{\epsilon_n})$,

$$\mathbb{S}^n \rightharpoonup \mathbb{S} \qquad \text{weakly in } L^{r'}(\Omega)^{3\times3}_{\mathrm{sym}}, \tag{3.29a}$$

$$\mathbb{D}\boldsymbol{v}^n \rightharpoonup \mathbb{D}\boldsymbol{v} \qquad \text{weakly in } L^{r}(\Omega)^{3\times3}_{\mathrm{sym}}, \tag{3.29b}$$

$$\boldsymbol{v}^n \to \boldsymbol{v} \qquad \text{strongly in } L^q(\Omega)^3 \text{ for all } q \in \left[1, \tfrac{3r}{3-r}\right). \tag{3.29c}$$

The last piece of information provides the strong convergence of $\{\boldsymbol{v}^n\}_{n=1}^\infty$ in $L^2(\Omega)^3$ provided that $\frac{3r}{3-r} > 2$, which gives the bound stated in the formulation of the theorem, namely $r > \frac{6}{5}$.

[9] In order to verify that

$$\limsup_{N\to\infty} \int_\Omega \mathbb{S}^N : \mathbb{D}\boldsymbol{v}^N \leq \int_\Omega \mathbb{S} : \mathbb{D}\boldsymbol{v},$$

one uses the identities

$$\int_\Omega \mathbb{S}^N : \mathbb{D}\boldsymbol{v}^N + \epsilon \int_\Omega \left|\mathbb{D}\boldsymbol{v}^N\right|^2 = \left\langle \boldsymbol{b}, \boldsymbol{v}^N \right\rangle,$$

$$\int_\Omega \mathbb{S} : \mathbb{D}\boldsymbol{v} + \epsilon \int_\Omega \left|\mathbb{D}\boldsymbol{v}\right|^2 = \left\langle \boldsymbol{b}, \boldsymbol{v} \right\rangle,$$

the weak lower semicontinuity of the L^2-norm of $\mathbb{D}\boldsymbol{v}^N$, and the inequality

$$\liminf_{n\to\infty} a_n + \limsup_{n\to\infty} b_n \leq \limsup_{n\to\infty} (a_n + b_n)$$

applied to any sequences $\{a_n\}_{n=1}^\infty$, $\{b_n\}_{n=1}^\infty$ with $a_n \geq 0$, $b_n \geq 0$ for all $n \in \mathbb{N}$.

Consequently, $\boldsymbol{v}$ and $\mathbb{S}$ fulfill (3.14). It remains to prove that $(\mathbb{S}, \mathbb{D}\boldsymbol{v}) \in \mathcal{G}$ a.e. in Ω. By the graph convergence lemma (Lemma A.6), it is enough to show that

$$\limsup_{n\to\infty} \big(\mathbb{S}^n, \mathbb{D}\boldsymbol{v}^n\big) \leq \big(\mathbb{S}, \mathbb{D}\boldsymbol{v}\big).$$

To prove it, we first subtract (3.14) from (3.27a) to obtain

$$\begin{aligned} (\mathbb{S}^n - \mathbb{S}, \mathbb{D}\boldsymbol{\varphi}) + \epsilon_n\, (\mathbb{D}\boldsymbol{v}^n, \mathbb{D}\boldsymbol{\varphi}) + (\boldsymbol{v}\otimes\boldsymbol{v} - \boldsymbol{v}^n \otimes \boldsymbol{v}^n, \mathbb{D}\boldsymbol{\varphi}) = 0& \\ \text{for all } \boldsymbol{\varphi} \in \mathbf{W}^{1,2}_{0,\mathrm{div}} \cap \mathbf{W}^{1,\frac{3r}{5r-6}}_{0,\mathrm{div}}.& \end{aligned} \tag{3.30}$$

Let $B \subset \mathbb{R}^3$ be an arbitrary ball of radius R such that $\frac{B}{2} \subset B \subset 2B \subset \Omega$ and $\chi \in \mathcal{C}^\infty_0(B)$ be such that $\chi = 1$ in $\frac{B}{2}$, $\chi \leq 1$ in B, and $|\nabla\chi| \leq CR^{-1}$ in B. Then we set

$$\boldsymbol{u}^n := \chi\big(\boldsymbol{v}^n - \boldsymbol{v}\big) - \boldsymbol{h}^n,$$

where $\boldsymbol{h}^n \in \mathbf{W}^{1,r}_0(B)$ solves

$$\operatorname{div} \boldsymbol{h}^n = \nabla\chi \cdot \big(\boldsymbol{v}^n - \boldsymbol{v}\big) \qquad \text{in } B,$$

which is solvable as the compatibility condition $\int_B \nabla\chi \cdot \big(\boldsymbol{v}^n - \boldsymbol{v}\big) = 0$ is met. In fact, there is a continuous linear operator $\mathcal{B} : \{q \in L^p(B), \int_\Omega q = 0\} \to \mathbf{W}^{1,p}_0(B) : g \mapsto \boldsymbol{u}$ such that $\operatorname{div} \boldsymbol{u} = g$, cf. Remark A.8 (ii). Consequently, $\boldsymbol{u}^n$ extended by zero in $\Omega \setminus B$ fulfill $\operatorname{div} \boldsymbol{u}^n = 0$ in Ω. Next we consider divergence-free Lipschitz approximations $\boldsymbol{u}^{n,k}$ to $\boldsymbol{u}^n$ from Lemma A.4. Taking $\boldsymbol{\varphi} := \boldsymbol{u}^{n,k}$ in (3.30) and letting $n \to \infty$, we conclude, using in particular the property (d) of Lemma A.4 and (3.28), that

$$\lim_{n\to\infty} \big(\mathbb{S}^n - \mathbb{S}, \mathbb{D}\boldsymbol{u}^{n,k}\big) = 0. \tag{3.31}$$

Using the properties of $\boldsymbol{u}^{n,k}$ (see Lemma A.4) and splitting the integral on the left-hand side of (3.31) into integrals over $\mathcal{O}^{n,k}$ and $B \setminus \mathcal{O}^{n,k}$, we conclude

$$\lim_{n\to\infty} \big(\mathbb{S}^n - \mathbb{S}, \mathbb{D}\boldsymbol{u}^n\big)_{B\setminus\mathcal{O}^{n,k}} \leq C2^{-k} \quad \text{for all } k \in \mathbb{N}.$$

It follows from the definition of $\boldsymbol{u}^n$, the properties of the operator $\mathcal{B}$ and the compactness of $\boldsymbol{v}^n$ that

$$\lim_{n\to\infty} \int_{B\setminus\mathcal{O}^{n,k}} (\mathbb{S}^n - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^n - \mathbb{D}\boldsymbol{v})\, \chi \leq C2^{-k} \quad \text{for arbitrary } k \in \mathbb{N},$$

which implies, by applying the Hölder inequality, that

$$\lim_{n\to\infty} \int_B \big|(\mathbb{S}^n - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^n - \mathbb{D}\boldsymbol{v})\big|^{\frac{1}{2}} \chi^{\frac{1}{2}} \leq C2^{-k} \quad \text{for arbitrary } k \in \mathbb{N}.$$

This leads to

$$\lim_{n\to\infty} \int_{\frac{B}{2}} \big|(\mathbb{S}^n - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^n - \mathbb{D}\boldsymbol{v})\big|^{\frac{1}{2}}\, \mathrm{d}x \leq C2^{-k} \quad \text{for arbitrary } k \in \mathbb{N}.$$

Let us set $g^n := \big|(\mathbb{S}^n - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^n - \mathbb{D}\boldsymbol{v})\big|$. Clearly $g^n \geq 0$ and $g^n \to 0$ almost everywhere in $\frac{B}{2}$. But as B is arbitrary, we conclude that

$$g^n \to 0 \text{ almost everywhere in } \Omega. \tag{3.32}$$

Since $\{g^n\}_{n=1}^\infty$ is bounded in $L^1(\Omega)$ and has the pointwise limit (3.32), Corollary A.3, a consequence of the biting lemma (Lemma A.2), then implies existence of a subsequence $\{g^{n_j}\}_{j=1}^\infty$, and a sequence of sets $\{E_k\}_{k=1}^\infty$ with $\Omega \supset E_1 \supset E_2 \supset \ldots$, $|E_k| \to 0$ such that for all $k \in \mathbb{N}$

$$g^{n_j} \to 0 \text{ strongly in } L^1(\Omega \setminus E_k).$$

From the definition of g^{n_j} we conclude that

$$\limsup_{j\to\infty} \int_{\Omega\setminus E_k} (\mathbb{S}^{n_j} - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^{n_j} - \mathbb{D}\boldsymbol{v}) = 0,$$

which implies as a consequence of (3.29a) and (3.29b) that for all $k \in \mathbb{N}$

$$\limsup_{j\to\infty} \int_{\Omega\setminus E_k} \mathbb{S}^{n_j} : \mathbb{D}\boldsymbol{v}^{n_j} = \int_{\Omega\setminus E_k} \mathbb{S} : \mathbb{D}\boldsymbol{v}.$$

Since $|E_k| \to 0$, we can conclude from the graph convergence lemma (Lemma A.6) that

$$(\mathbb{S}, \mathbb{D}\boldsymbol{v}) \in \mathcal{G} \text{ almost everywhere in } \Omega$$

so that (3.13) holds and the first part of the theorem is proved.

On the pressure. Setting

$$\langle \boldsymbol{F}, \boldsymbol{\varphi}\rangle := (\mathbb{S}, \mathbb{D}\boldsymbol{\varphi}) - (\boldsymbol{v}\otimes\boldsymbol{v}, \nabla\boldsymbol{\varphi}) - \langle \boldsymbol{b}, \boldsymbol{\varphi}\rangle$$

we observe that

$$\langle \boldsymbol{F}, \boldsymbol{\varphi}\rangle = 0 \quad \text{for all } \boldsymbol{\varphi} \in \boldsymbol{\mathcal{C}}^{\infty}_{0,\mathrm{div}}$$

and

$$\boldsymbol{F} \in \begin{cases} \left(\mathbf{W}^{1,r}_0\right)^* & \text{if } r \geq \frac{9}{5}, \\ \left(\mathbf{W}^{1,\frac{3r}{5r-6}}_0\right)^* & \text{if } r \in \left(\frac{6}{5}, \frac{9}{5}\right). \end{cases} \tag{3.33}$$

By the de Rham theorem, see [6, Theorem 2.1], there is $p \in \left(\mathcal{C}^{\infty}_0(\Omega)\right)^*$ such that

$$\langle \boldsymbol{F}, \boldsymbol{\varphi}\rangle = \langle -\nabla p, \boldsymbol{\varphi}\rangle \quad \text{for all } \boldsymbol{\varphi} \in \boldsymbol{\mathcal{C}}^{\infty}_0. \tag{3.34}$$

Since Ω is $C^{0,1}$ domain, the Nečas theorem (see Lemma A.7, Remark A.9) together with (3.33) and (3.34) implies (3.15) and (3.16). □

3.2.2 Slip case

In this part we replace the no-slip boundary condition either by

$$\boldsymbol{v}\cdot\boldsymbol{n} = 0 \quad \text{and} \quad \boldsymbol{s} = \boldsymbol{0} \quad \text{on } \partial\Omega \tag{3.35}$$

or by

$$\begin{gathered} \boldsymbol{v}\cdot\boldsymbol{n} = 0 \quad \text{and} \quad (\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) \in \mathcal{B} \quad \text{on } \partial\Omega, \\ \text{where } \mathcal{B} \text{ fulfills the conditions } (\mathcal{B}1)\text{–}(\mathcal{B}4). \end{gathered} \tag{3.36}$$

We prove the following result.

Theorem 3.2. *Let $\Omega \subset \mathbb{R}^3$ be a $C^{1,1}$ domain.*[10] *Let further $r > \frac{6}{5}$, $\boldsymbol{b} \in \left(\mathbf{W}^{1,r}_{\boldsymbol{n}}\right)^*$, $\mathcal{G}$ be a maximal monotone r-graph of the form* (3.10).

(i) (Boundary condition (3.36)) *Let $\mathcal{B}$ be a maximal monotone* 2*-graph. Then there is a weak solution*

$$(\boldsymbol{v}, \mathbb{S}, \boldsymbol{s}) \in \mathbf{W}^{1,r}_{\boldsymbol{n},\mathrm{div}} \times L^{r'}(\Omega)^{3\times3}_{\mathrm{sym}} \times L^2(\partial\Omega)^3$$

to (3.9) *and* (3.36) *such that*

$$\big(\mathbb{S}, \mathbb{D}\boldsymbol{v}\big) \in \mathcal{G} \text{ a.e. in } \Omega, \tag{3.37}$$

$$\big(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}\big) \in \mathcal{B} \text{ a.e. in } \partial\Omega, \tag{3.38}$$

and

$$\big(\mathbb{S}, \mathbb{D}\boldsymbol{\varphi}\big) - \big(\boldsymbol{v}\otimes\boldsymbol{v}, \mathbb{D}\boldsymbol{\varphi}\big) + \big(\boldsymbol{s}, \boldsymbol{\varphi}\big)_{\partial\Omega} = \big\langle \boldsymbol{b}, \boldsymbol{\varphi}\big\rangle$$

$$\text{for all } \boldsymbol{\varphi} \in \begin{cases} \mathbf{W}^{1,r}_{\boldsymbol{n},\mathrm{div}} & \text{if } r \geq \frac{9}{5}, \\ \mathbf{W}^{1,\frac{3r}{5r-6}}_{\boldsymbol{n},\mathrm{div}} & \text{if } r \in \left(\frac{6}{5}, \frac{9}{5}\right). \end{cases} \tag{3.39}$$

[10] In the case of the boundary condition (3.35), the required regularity of the boundary can be weakened at the cost of losing the information concerning the pressure.

(ii) (Boundary condition (3.35)) *Assume that* Ω *is not axisymmetric. Then there is a weak solution*

$$(\boldsymbol{v},\mathbb{S}) \in \mathbf{W}^{1,r}_{\boldsymbol{n},\mathrm{div}} \times L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}$$

to (3.9) *and* (3.35) *such that* (3.37) *and* (3.39) *with* $\boldsymbol{s}=\boldsymbol{0}$ *hold true.*

In addition, there is

$$p \in \begin{cases} L^{r'}(\Omega) & \textit{if } r \geq \frac{9}{5} \\ L^{\frac{3r}{2(3-r)}}(\Omega) & \textit{if } r \in \left(\frac{6}{5},\frac{9}{5}\right) \end{cases}, \qquad \int_\Omega p \,\mathrm{d}x = 0 \tag{3.40}$$

such that

$$\big(\mathbb{S},\mathbb{D}\boldsymbol{\varphi}\big) - \big(\boldsymbol{v}\otimes\boldsymbol{v},\mathbb{D}\boldsymbol{\varphi}\big) + \big(\boldsymbol{s},\boldsymbol{\varphi}\big)_{\partial\Omega} = \big(p,\operatorname{div}\boldsymbol{\varphi}\big) + \big\langle \boldsymbol{b},\boldsymbol{\varphi}\big\rangle$$
$$\textit{for all } \boldsymbol{\varphi} \in \begin{cases} \mathbf{W}^{1,r}_{\boldsymbol{n}} & \textit{if } r \geq \frac{9}{5}, \\ \mathbf{W}^{1,\frac{3r}{5r-6}}_{\boldsymbol{n}} & \textit{if } r \in \left(\frac{6}{5},\frac{9}{5}\right). \end{cases} \tag{3.41}$$

Proof. The proof of existence of $\boldsymbol{v}$, $\mathbb{S}$, $\boldsymbol{s}$ with (3.37), (3.38), and (3.39) follows the same scheme as in the case of the no-slip boundary condition. The only differences are

(i) due to a different choice of the function spaces for the velocity as $\mathbf{W}^{1,r}_{0,\mathrm{div}}$ is replaced by $\mathbf{W}^{1,r}_{\boldsymbol{n},\mathrm{div}}$,

(ii) due to the presence of the term $\int_{\partial\Omega} \boldsymbol{s}\cdot\boldsymbol{\varphi}$ in the weak formulation of the balance of linear momentum,

(iii) and due to the necessity to verify validity of the boundary condition $\boldsymbol{h}(\boldsymbol{s},\boldsymbol{v_\tau}) = \boldsymbol{0}$.

Note that in the case $\boldsymbol{s}=\boldsymbol{0}$ on $\partial\Omega$, the last two differences disappear. In the case of the stick/slip boundary condition $\boldsymbol{v_\tau} = \frac{1}{\gamma_*}\frac{(|\boldsymbol{s}|-\sigma_*)^+}{|\boldsymbol{s}|}\boldsymbol{s}$ we start with its approximation

$$\boldsymbol{v_\tau} = \frac{1}{\gamma_*}\frac{(|\boldsymbol{s}|-\sigma_*)^+}{|\boldsymbol{s}|}\boldsymbol{s} + \frac{\epsilon}{\gamma_*}\boldsymbol{s}, \quad \epsilon > 0.$$

Other boundary conditions of the type (3.36) employ similar approximations.

For $r \geq \frac{9}{5}$, the proof then proceeds as in the case of no-slip boundary conditions up to the use of the appropriate form of Korn's inequality ((3.6) for boundary condition (3.35) assuming the domain is not axisymmetric, or (3.5) for boundary condition (3.36)) and the following points. The additional term $\int_{\partial\Omega} \boldsymbol{s}^N\cdot\boldsymbol{\varphi}$ in the balance of linear momentum is treated using the weak convergence $\boldsymbol{s}^N \rightharpoonup \boldsymbol{s}$ in $L^2(\partial\Omega)$. Since $W^{1,r}(\Omega) \hookrightarrow W^{1-\frac{1}{r},r}(\partial\Omega) \hookrightarrow\hookrightarrow L^2(\partial\Omega)$ provided $r > \frac{3}{2}$ and Ω is a $C^{0,1}$ domain, the space $\mathbf{W}^{1,r}_{\boldsymbol{n}}$ is compactly embedded into $L^2(\partial\Omega)$ even for $r > \frac{3}{2}$, and consequently $\boldsymbol{v}^N_{\boldsymbol{\tau}} \to \boldsymbol{v_\tau}$ strongly in $L^2(\partial\Omega)$ and thus

$$\int_{\partial\Omega} \boldsymbol{s}^N\cdot\boldsymbol{v}^N_{\boldsymbol{\tau}} \to \int_{\partial\Omega} \boldsymbol{s}\cdot\boldsymbol{v_\tau}. \tag{3.42}$$

This implies (see Lemma A.6) that $(\boldsymbol{s},\boldsymbol{v_\tau}) \in \mathcal{B}$ a.e. on $\partial\Omega$.

For $r \in \left(\frac{6}{5},\frac{9}{5}\right)$, the validity of $(\mathbb{S},\mathbb{D}\boldsymbol{v}) \in \mathcal{G}$ a.e. in Ω can be established in the same way as in the no-slip case since the proof is based on local (interior) arguments. It remains to show that $(\boldsymbol{s},\boldsymbol{v_\tau}) \in \mathcal{B}$ a.e. on $\partial\Omega$. For $r > \frac{3}{2}$, it follows from (3.42) and Lemma A.6. To use Lemma A.6 also for $r \in \left(\frac{6}{5},\frac{3}{2}\right]$, it suffices to show that

$$\limsup_{N\to\infty} \int_{\partial\Omega} \boldsymbol{s}^N\cdot\boldsymbol{v}^N \leq \int_{\partial\Omega} \boldsymbol{s}\cdot\boldsymbol{v}.$$

However, for $r > 1$: $W^{1,r}(\Omega) \hookrightarrow\hookrightarrow L^1(\partial\Omega)$, the strong convergence $\boldsymbol{v}^N_{\boldsymbol{\tau}} \to \boldsymbol{v_\tau}$ in $L^1(\partial\Omega)$ together with Egorov's theorem implies that for any $\delta > 0$ there is $\mathcal{U}_\delta \subset \partial\Omega$ such that $|\partial\Omega\setminus\mathcal{U}_\delta| < \delta$ and $\boldsymbol{v}^N_{\boldsymbol{\tau}} \to \boldsymbol{v_\tau}$ strongly in $L^\infty(\mathcal{U}_\delta)^3$. Hence

$$\limsup_{N\to\infty} \int_{\mathcal{U}_\delta} \boldsymbol{s}^N\cdot\boldsymbol{v}^N_{\boldsymbol{\tau}} \leq \int_{\mathcal{U}_\delta} \boldsymbol{s}\cdot\boldsymbol{v_\tau}.$$

Consequently, by the graph convergence lemma (Lemma A.6 in Appendix), $(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) \in \mathcal{B}$ a.e. in $\mathcal{U}_\delta$. As $\delta > 0$ was arbitrary, we conclude that $(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) \in \mathcal{B}$ a.e. on $\partial\Omega$. The proof of the first part of the theorem is complete.

It remains to prove the existence of pressure (3.40) fulfilling (3.41). Let us define a linear functional $\boldsymbol{F}$ through the relation

$$\big\langle \boldsymbol{F}, \boldsymbol{\varphi} \big\rangle := \big\langle \boldsymbol{b}, \boldsymbol{\varphi} \big\rangle - \big(\mathbb{S} - \boldsymbol{v} \otimes \boldsymbol{v}, \mathbb{D}\boldsymbol{\varphi} \big) - \big(\boldsymbol{s}, \boldsymbol{\varphi} \big)_{\partial\Omega} \quad \text{for any } \boldsymbol{\varphi} \in \mathcal{C}^\infty_{\boldsymbol{n}}.$$

From (3.39) we can see that $\boldsymbol{F} \in \big(\mathbf{W}^{1,q}_{\boldsymbol{n}}\big)^*$ where $q = r$ if $r \geq \frac{9}{5}$ and $q = \frac{3r}{5r-6}$ if $r \in (\frac{6}{5}, \frac{9}{5})$ and

$$\big\langle \boldsymbol{F}, \boldsymbol{\varphi} \big\rangle = 0 \quad \text{for all } \boldsymbol{\varphi} \in \mathbf{W}^{1,q}_{\boldsymbol{n},\text{div}}. \tag{3.43}$$

Now consider a variational problem to find $p \in L^{q'}(\Omega)$ with $\int_\Omega p = 0$ such that

$$\big(p, -\Delta\phi\big) = \big\langle \boldsymbol{F}, \nabla\phi \big\rangle \quad \text{for all } \phi \in W^{2,q}(\Omega) \text{ with } \nabla\phi \in \mathbf{W}^{1,q}_{\boldsymbol{n}}. \tag{3.44}$$

As a consequence of the $C^{1,1}$ smoothness of the domain, one can employ Lemma A.11 to conclude that (3.44) is equivalent with the problem: find $p \in L^{q'}(\Omega)$ such that $\int_\Omega p = 0$ and

$$\big(p, q\big) = \big\langle \boldsymbol{F}, \nabla A^{-1} q \big\rangle \quad \text{for all } q \in L^q(\Omega) \text{ with } \int_\Omega q = 0 \tag{3.45}$$

where A^{-1} is the solution operator for the Neumann-Poisson problem (A.2). The problem (3.45) has a unique solution by virtue of Lemma A.11. Thus we have constructed p with properties (3.40). To verify (3.41), consider a test function $\boldsymbol{\varphi} \in \mathbf{W}^{1,q}_{\boldsymbol{n}}$. With the Helmholtz decomposition (see Corollary A.12 in Appendix) $\boldsymbol{\varphi} = \nabla\phi + \boldsymbol{\varphi}_0$ with $\nabla\phi \in \mathbf{W}^{1,q}_{\boldsymbol{n}}$ and $\boldsymbol{\varphi}_0 \in \mathbf{W}^{1,q}_{\boldsymbol{n},\text{div}}$ we can immediately obtain, using (3.43) and (3.44), that

$$\big\langle \boldsymbol{F}, \boldsymbol{\varphi} \big\rangle = \big\langle \boldsymbol{F}, \nabla\phi \big\rangle + \big\langle \boldsymbol{F}, \boldsymbol{\varphi}_0 \big\rangle = \big(p, -\Delta\phi\big) = -\big(p, \operatorname{div}\boldsymbol{\varphi}\big).$$

This proves (3.41). The proof of Theorem 3.2 is thus complete. □

3.3 Analysis of unsteady flows

In this section, we investigate unsteady internal flows, i.e., flows governed by (3.1). Again, we treat separately two cases: the no-slip boundary condition and the boundary conditions allowing slip.

3.3.1 No-slip case

We first provide an existence result for the no-slip case, i.e., we investigate the system (3.1a)–(3.1d), (3.1f), and $\boldsymbol{v}_{\boldsymbol{\tau}} = \mathbf{0}$ on $(0,T) \times \partial\Omega$ as a special case of (3.1e).

Theorem 3.3. *Let $T \in (0,\infty)$, $\Omega \subset \mathbb{R}^3$ be a domain and $Q := (0,T) \times \Omega$. Let $r > \frac{6}{5}$, $\boldsymbol{b} \in L^{r'}\big(0,T; (\mathbf{W}^{1,r}_0)^*\big)$ and $\boldsymbol{v}_0 \in \mathbf{L}^2_{\boldsymbol{n},\text{div}}$. Let further $\mathcal{G} \subset \mathbb{R}^{3\times3}_{\text{sym}} \times \mathbb{R}^{3\times3}_{\text{sym}}$ be a maximal monotone r-graph of the form* (3.10) *fulfilling (G1)–(G4). Then there exists a pair $(\boldsymbol{v}, \mathbb{S})$:*

$$\boldsymbol{v} \in L^\infty(0,T; \mathbf{L}^2_{\boldsymbol{n},\text{div}}) \cap L^r(0,T; \mathbf{W}^{1,r}_{0,\text{div}}), \tag{3.46a}$$

$$\mathbb{S} \in L^{r'}(Q)^{3\times3}_{\text{sym}} \tag{3.46b}$$

satisfying

$$\lim_{t\to0+} \int_\Omega |\boldsymbol{v}(t,\cdot) - \boldsymbol{v}_0|^2 = 0, \tag{3.46c}$$

$$\begin{gathered} \int_Q \mathbb{S} : \mathbb{D}\boldsymbol{w} = \int_0^T \langle \boldsymbol{b}, \boldsymbol{w} \rangle + \int_Q \boldsymbol{v} \otimes \boldsymbol{v} : \mathbb{D}\boldsymbol{w} + \int_Q \boldsymbol{v} \cdot \frac{\partial \boldsymbol{w}}{\partial t} + \int_\Omega \boldsymbol{v}_0 \cdot \boldsymbol{w}(0,\cdot) \\ \textit{for all } \boldsymbol{w} \in \mathcal{C}^\infty_0\big([0,T); \mathbf{W}^{1,q}_{0,\text{div}}\big),\ q = \max\{r, \tfrac{5r}{5r-6}\}, \end{gathered} \tag{3.46d}$$

$$\mathbb{S} = 2\nu_* \left(|\mathbb{D}\boldsymbol{v}| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}\boldsymbol{v}|) \tfrac{\mathbb{D}\boldsymbol{v}}{|\mathbb{D}\boldsymbol{v}|} \quad \textit{almost everywhere in } Q. \tag{3.46e}$$

Moreover, the energy inequality holds:

$$\int_\Omega \frac{|\boldsymbol{v}(t,\cdot)|^2}{2} + \int_0^t \int_\Omega \mathbb{S} : \mathbb{D}\boldsymbol{v} \le \int_\Omega \frac{|\boldsymbol{v}_0|^2}{2} + \int_0^t \langle \boldsymbol{b}, \boldsymbol{v}\rangle \qquad \text{for almost all } t \in (0,T) \text{ and for } t = T; \tag{3.47}$$

if $r \ge \frac{11}{5}$, (3.47) *becomes equality.*

In addition, if Ω is a $C^{0,1}$ domain with sufficiently small Lipschitz constant (smallness depending only on r) or Ω is any C^1 domain, then there are $P^1 \in L^\infty(0,T;L^6(\Omega))$, $P^1(t,\cdot)$ harmonic in Ω for almost every $t \in (0,T)$ and $p^2 \in L^{q'}(Q)$ with $q = \max\{r, \frac{5r}{5r-6}\}$ such that

$$\begin{aligned} \int_Q \mathbb{S} : \mathbb{D}\boldsymbol{\omega} = \int_0^T \langle \boldsymbol{b}, \boldsymbol{\omega}\rangle + \int_Q \boldsymbol{v}\otimes\boldsymbol{v} : \mathbb{D}\boldsymbol{\omega} + \int_Q \boldsymbol{v}\cdot\frac{\partial\boldsymbol{\omega}}{\partial t} + \int_\Omega \boldsymbol{v}_0\cdot\boldsymbol{\omega}(0,\cdot) \\ - \int_Q P^1 \operatorname{div}\frac{\partial\boldsymbol{\omega}}{\partial t} + \int_Q p^2 \operatorname{div}\boldsymbol{\omega} \qquad \text{for all } \boldsymbol{\omega} \in \mathcal{C}_0^\infty\big([0,T);\mathbf{W}_0^{1,q}\big), \\ q = \max\{r, \tfrac{5r}{5r-6}\}. \end{aligned} \tag{3.48}$$

Functions P^1 and p^2 can be chosen such that $\int_\Omega P^1(t,\cdot) = \int_\Omega p^2(t,\cdot) = 0$ for almost every $t \in (0,T)$. If, in addition, Ω is a $C^{1,1}$ domain then it holds $\nabla P^1 \in L^\infty(0,T;L^2(\Omega)) \cap L^{\frac{5r}{3}}(Q)$.

Remark 3.4. *(1) We could define the weak solution to the problem considered differently. We could say that $\boldsymbol{v}$ is a weak solution to the problem if $\boldsymbol{v}$ fulfills* (3.46a), (3.46c) *and*

$$\begin{aligned} \int_Q 2\nu_* \left(|\mathbb{D}\boldsymbol{v}| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}\boldsymbol{v}|)\tfrac{\mathbb{D}\boldsymbol{v}}{|\mathbb{D}\boldsymbol{v}|} : \mathbb{D}\boldsymbol{w} = \int_0^T \langle \boldsymbol{b}, \boldsymbol{w}\rangle + \int_Q \boldsymbol{v}\otimes\boldsymbol{v} : \mathbb{D}\boldsymbol{w} \\ + \int_Q \boldsymbol{v}\cdot\frac{\partial\boldsymbol{w}}{\partial t} + \int_\Omega \boldsymbol{v}_0\cdot\boldsymbol{w}(0,\cdot) \\ \text{for all } \boldsymbol{w} \in \mathcal{C}_0^\infty\big([0,T);\mathbf{W}_{0,\mathrm{div}}^{1,q}\big),\ q = \max\{r, \tfrac{5r}{5r-6}\}. \end{aligned}$$

(2) It holds $\boldsymbol{v} \in \mathcal{C}\big(0,T;L^2(\Omega)^3\big)$ if $r \ge \frac{11}{5}$ and $\boldsymbol{v} \in \mathcal{C}\big(0,T;L^2_{\mathrm{weak}}(\Omega)^3\big)$ if $r \in (\frac{6}{5}, \frac{11}{5})$.

Proof of Theorem 3.3. We shall distinguish two cases (that can be also identified via behavior of the total dissipation of energy with respect to scaling invariance of the governing equations, see [58]): the subcritical/critical case $r \ge \frac{11}{5}$ and the supercritical case $r \in \left(\frac{6}{5}, \frac{11}{5}\right)$. The problem can be analyzed in an arbitrary spatial dimension d; then the supercritical case corresponds to $r \in \left(\frac{2d}{d+2}, 1 + \frac{2d}{d+2}\right)$ and the subcritical/critical case to $r \ge 1 + \frac{2d}{d+2}$. Note that the case $r = 2$ (including the Euler/Navier-Stokes fluid) belongs to the supercritical case in any spatial dimension $d > 2$.

The case $r \ge \frac{11}{5}$. Step 1. Galerkin approximations. We first construct a finite-dimensional approximation to the problem by the Galerkin method. To proceed, we consider an auxiliary eigenvalue problem to find $\lambda \in \mathbb{R}$ and $\boldsymbol{\omega} \in \mathbf{W}^{3,2}_{0,\mathrm{div}} \hookrightarrow W^{1,\infty}(\Omega)^3$ satisfying

$$(\!(\boldsymbol{\omega}, \boldsymbol{\varphi})\!) = \lambda(\boldsymbol{\omega}, \boldsymbol{\varphi}) \text{ for all } \boldsymbol{\varphi} \in \mathbf{W}^{3,2}_{0,\mathrm{div}}, \tag{3.49}$$

where $(\cdot,\cdot)$ is a scalar product in $L^2(\Omega)^3$ and $(\!(\cdot,\cdot)\!)$ is a scalar product in $\mathbf{W}^{3,2}_{0,\mathrm{div}}$, i.e., $(\!(\boldsymbol{\omega}, \boldsymbol{\varphi})\!) := (\nabla^3\boldsymbol{\omega}, \nabla^3\boldsymbol{\varphi}) + (\boldsymbol{\omega}, \boldsymbol{\varphi})$. It is known, see for example [56, Appendix A.4], that there exist eigenvalues $\{\lambda_m\}_{m=1}^\infty$ and corresponding eigenfunctions $\{\boldsymbol{\omega}^m\}_{m=1}^\infty$ for the eigenvalue problem (3.49) such that $0 < \lambda_1 \le \lambda_2 \le \ldots$, $\lambda_m \to \infty$ as $m \to \infty$, $(\boldsymbol{\omega}^m, \boldsymbol{\omega}^n) = \delta_{mn}$, $\left(\!\!\left(\frac{\boldsymbol{\omega}^m}{\sqrt{\lambda_m}}, \frac{\boldsymbol{\omega}^n}{\sqrt{\lambda_n}}\right)\!\!\right) = \delta_{mn}$. Furthermore, the mappings $\boldsymbol{P}^N : \mathbf{W}^{3,2}_{0,\mathrm{div}} \to H^N := \operatorname{span}\{\boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \ldots, \boldsymbol{\omega}^N\}$ defined by $\boldsymbol{P}^N\boldsymbol{v} := \sum_{i=1}^N (\boldsymbol{v}, \boldsymbol{\omega}^i)\,\boldsymbol{\omega}^i$ are continuous orthonormal projectors in $L^2(\Omega)^3$, $\mathbf{W}^{3,2}_{0,\mathrm{div}}$ and $\big(\mathbf{W}^{3,2}_{0,\mathrm{div}}\big)^*$, in particular

$$\|\boldsymbol{P}^N\|_{\mathcal{L}(L^2(\Omega)^3)} \le 1, \qquad \|\boldsymbol{P}^N\|_{\mathcal{L}\left(\mathbf{W}^{3,2}_{0,\mathrm{div}}\right)} \le 1, \qquad \|\boldsymbol{P}^N\|_{\mathcal{L}\left(\left(\mathbf{W}^{3,2}_{0,\mathrm{div}}\right)^*\right)} \le 1. \tag{3.50}$$

Galerkin approximations $\boldsymbol{v}^N(t) \in H^N$ of the form $\boldsymbol{v}^N(t,x) = \sum_{j=1}^N c_j^N(t)\,\boldsymbol{\omega}^j(x)$ are introduced in such a way that the coefficients $\boldsymbol{c}^N = (c_1^N, c_2^N, \dots, c_N^N)$ fulfill

$$\begin{aligned}\left(\frac{\mathrm{d}\boldsymbol{v}^N}{\mathrm{d}t}, \boldsymbol{\omega}^j\right) - \big(\boldsymbol{v}^N \otimes \boldsymbol{v}^N, \nabla \boldsymbol{\omega}^j\big) + \big(\mathbb{S}(\mathbb{D}\boldsymbol{v}^N), \mathbb{D}\boldsymbol{\omega}^j\big) = \big(\boldsymbol{P}^N \boldsymbol{b}, \boldsymbol{\omega}^j\big) \quad j = 1,2,\dots,N,\\ \boldsymbol{v}^N(0,\cdot) = \boldsymbol{P}^N \boldsymbol{v}_0,\end{aligned} \tag{3.51}$$

where

$$\mathbb{S}(\mathbb{D}) := 2\nu_* \left(|\mathbb{D}| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}|) \tfrac{\mathbb{D}}{|\mathbb{D}|}.$$

Since the mappings $\boldsymbol{z} \mapsto \boldsymbol{z} \otimes \boldsymbol{z}$ and $\boldsymbol{z} \mapsto \mathbb{S}(\mathbb{D}\boldsymbol{z})$ are continuous, the Carathéodory theory for systems of ordinary differential equations implies local existence of a solution $\boldsymbol{c}^N$ solving (3.51). Global existence then follows from the fact that

$$\sup_{t \in (0,T)} |\boldsymbol{c}^N(t)|_{\mathbb{R}^N} < \infty.$$

This piece of information is a simple consequence of the orthogonality of the basis $\{\boldsymbol{\omega}^j\}_{j=1}^\infty$ and a priori estimates that will follow, see (3.54) below.

Step 2. Uniform estimates and their consequences. Multiplying (3.51) by $c_j^N(t)$, taking the sum over $j = 1, 2, \dots, N$, using the fact $(\boldsymbol{z} \otimes \boldsymbol{z}, \nabla \boldsymbol{z}) = 0$ for $\boldsymbol{z}$ with $\operatorname{div} \boldsymbol{z} = 0$, $\boldsymbol{z} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}^N\|_2^2 + (\mathbb{S}(\mathbb{D}\boldsymbol{v}^N), \mathbb{D}\boldsymbol{v}^N) = (\boldsymbol{P}^N \boldsymbol{b}, \boldsymbol{v}^N). \tag{3.52}$$

Since $\mathcal{G}$ is an r-graph fulfilling ($\mathcal{G}4$), we conclude that for all $t \in (0,T]$

$$\|\boldsymbol{v}^N(t)\|_2^2 + \alpha \int_0^t \left(\|\mathbb{S}^N\|_{r'}^{r'} + \|\mathbb{D}\boldsymbol{v}^N\|_r^r\right) \le C\left(\beta, \|\boldsymbol{v}_0\|_2^2, \|\boldsymbol{b}\|_{(L^r(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}}))^*}\right). \tag{3.53}$$

Using the orthogonality of $\{\boldsymbol{\omega}^j\}_{j=1}^N$ in $L^2(\Omega)^3$, this, in particular, implies that

$$\sup_{t \in (0,T)} |\boldsymbol{c}^N(t)|_{\mathbb{R}^N} < \infty \tag{3.54}$$

so that the proof of global-in-time existence of $\boldsymbol{c}^N : [0,T] \to \mathbb{R}^N$ is complete.

Furthermor, Korn's inequality, see (3.2), the interpolation inequality

$$\|\boldsymbol{u}\|_q^q \le \|\boldsymbol{u}\|_2^{(1-\lambda)q} \|\boldsymbol{u}\|_{\frac{3r}{3-r}}^{\lambda q} \qquad \text{with } \lambda q = \frac{3r(q-2)}{5r-6},$$

and the embedding $\mathbf{W}_0^{1,r} \hookrightarrow L^{\frac{3r}{3-r}}(\Omega)^3$ together with (3.53) imply

$$\int_0^T \|\boldsymbol{v}^N\|_{\frac{5r}{3}}^{\frac{5r}{3}} \le C\left(\beta, \|\boldsymbol{v}_0\|_2^2, \|\boldsymbol{b}\|_{(L^r(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}}))^*}\right).$$

Finally, since for all $\boldsymbol{\varphi} \in L^s(0,T;\mathbf{W}^{3,2}_{0,\mathrm{div}})$

$$\int_0^T \left(\frac{\mathrm{d}\boldsymbol{v}^N}{dt}, \boldsymbol{\varphi}\right) = \int_0^T \left(\frac{\mathrm{d}\boldsymbol{v}^N}{dt}, \mathbb{P}^N \boldsymbol{\varphi}\right),$$

it follows from (3.51), (3.53), the fact that $2r' = \frac{2r}{r-1} \le \frac{5r}{3} \Leftrightarrow r \ge \frac{11}{5}$, and (3.50), that

$$\begin{aligned}\left\|\frac{\mathrm{d}\boldsymbol{v}^N}{\mathrm{d}t}\right\|_{\left(L^r\left(0,T;\mathbf{W}^{3,2}_{0,\mathrm{div}}\right)\right)^*} &:= \sup_{\substack{\boldsymbol{\varphi} \in L^r\left(0,T;\mathbf{W}^{3,2}_{0,\mathrm{div}}\right)\\ \|\boldsymbol{\varphi}\|_{L^r\left(0,T;\mathbf{W}^{3,2}_{0,\mathrm{div}}\right)} \le 1}} \left|\int_0^T \left(\frac{\mathrm{d}\boldsymbol{v}^N}{\mathrm{d}t}, \boldsymbol{\varphi}\right)\right|\\ &\le C\left(\beta, \|\boldsymbol{v}_0\|_2^2, \|\boldsymbol{b}\|_{\left(L^r\left(\mathbf{W}^{1,r}_{0,\mathrm{div}}\right)\right)^*}\right).\end{aligned}$$

Consequently, there are (not relabeled) subsequences so that

$$\begin{aligned}
\boldsymbol{v}^N &\overset{*}{\rightharpoonup} \boldsymbol{v} && *\text{-weakly in } L^\infty(0,T;L^2(\Omega)^3),\\
\mathbb{D}\boldsymbol{v}^N &\rightharpoonup \mathbb{D}\boldsymbol{v} && \text{weakly in } L^r(0,T;L^r(\Omega)^{3\times 3}_{\mathrm{sym}}),\\
\nabla\boldsymbol{v}^N &\rightharpoonup \nabla\boldsymbol{v} && \text{weakly in } L^r(0,T;L^r(\Omega)^{3\times 3}),\\
\mathbb{S}^N &\rightharpoonup \mathbb{S} && \text{weakly in } L^{r'}(0,T;L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}),\\
\partial_t\boldsymbol{v}^N &\rightharpoonup \partial_t\boldsymbol{v} && \text{weakly in } \big(L^r(0,T;\mathbf{W}^{3,2}_{0,\mathrm{div}})\big)^*,\\
\boldsymbol{v}^N &\to \boldsymbol{v} && \text{strongly in } L^q(0,T;L^q(\Omega)^3) \text{ for all } q\in\big[1,\tfrac{5r}{3}\big),
\end{aligned}\tag{3.55}$$

where the last limit (3.55) follows from the Aubin-Lions compactness lemma applied to $\mathbf{W}^{3,2}_{0,\mathrm{div}} \hookrightarrow \mathbf{W}^{1,r}_{0,\mathrm{div}} \hookrightarrow\hookrightarrow \mathbf{L}^r_{\boldsymbol{n},\mathrm{div}} \hookrightarrow \big(\mathbf{W}^{3,2}_{0,\mathrm{div}}\big)^*$.

Finally, letting $N\to\infty$ in (3.51) for $j\in\mathbb{N}$ arbitrary but fixed, one concludes that $(\boldsymbol{v},\mathbb{S})$ satisfy

$$\int_0^T \left(\left\langle \frac{\partial\boldsymbol{v}}{\partial t},\boldsymbol{\omega}\right\rangle - (\boldsymbol{v}\otimes\boldsymbol{v},\nabla\boldsymbol{\omega}) + (\mathbb{S},\mathbb{D}\boldsymbol{\omega})\right)\phi(t)\mathrm{d}t = \langle \boldsymbol{b},\boldsymbol{\omega}\phi\rangle$$

valid for all $\phi\in\mathcal{C}^\infty_0(-\infty,\infty)$ and $\boldsymbol{\omega}\in\mathbf{W}^{3,2}_{0,\mathrm{div}}$. Since the space $\mathbf{W}^{3,2}_{0,\mathrm{div}}$ is dense in $\mathbf{W}^{1,r}_{0,\mathrm{div}}$ as $r\geq\frac{11}{5}$, and functions of the form $\phi(t)\boldsymbol{\omega}(x)$ are dense in $L^r(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}})$, and finally

$$\begin{aligned}
\left|\int_0^T (\boldsymbol{v}\otimes\boldsymbol{v},\nabla\psi)\right| &\leq \left(\int_0^T \|\boldsymbol{v}\|_{2r'}^{2r'}\right)^{\frac{1}{r'}}\left(\int_0^T \|\nabla\psi\|_r^r\right)\\
&\leq C\left(\int_0^T \|\boldsymbol{v}\|_{\frac{5r}{3}}^{\frac{5r}{3}}\right)^{\frac{1}{r'}}\left(\int_0^T \|\nabla\psi\|_r^r\right) < +\infty,
\end{aligned}$$

we deduce that $(\boldsymbol{v},\mathbb{S})$ satisfies

$$\int_0^T \langle\partial_t\boldsymbol{v},\boldsymbol{\omega}\rangle + \int_Q \mathbb{S}:\mathbb{D}\boldsymbol{\omega} = \langle\boldsymbol{b},\boldsymbol{\omega}\rangle + \int_Q(\boldsymbol{v}\otimes\boldsymbol{v},\mathbb{D}\boldsymbol{\omega})$$
$$\text{for all } \boldsymbol{\omega}\in L^r(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}}).\tag{3.56}$$

This implies that $\partial_t\boldsymbol{v}\in L^{r'}\big(0,T;(\mathbf{W}^{1,r}_{0,\mathrm{div}})^*\big)$. Inserting $\boldsymbol{\omega}:=\boldsymbol{v}$ into (3.56), we obtain the energy equality (3.47). It remains to show (3.46e).

Step 3. Attainment of the constitutive equation. To prove (3.46e), we wish to use the graph convergence lemma (see Lemma A.6 in Appendix). To apply this lemma, we need to show that

$$\limsup_{N\to\infty}\int_Q \mathbb{S}^N:\mathbb{D}\boldsymbol{v}^N \leq \int_Q \mathbb{S}:\mathbb{D}\boldsymbol{v}.$$

However (3.55) implies, in particular, that

$$\boldsymbol{v}^N(t)\to\boldsymbol{v}(t) \quad \text{in } L^2(\Omega)^3 \text{ for almost all } t\in(0,T].$$

Integrating (3.52) from 0 to such t's, and letting $N\to\infty$, one concludes

$$\tfrac{1}{2}\|\boldsymbol{v}(t)\|_2^2 + \limsup_{N\to\infty}\int_0^t\int_\Omega \mathbb{S}^N:\mathbb{D}\boldsymbol{v}^N = \langle\boldsymbol{b},\boldsymbol{v}\rangle + \tfrac{1}{2}\|\boldsymbol{v}_0\|_2^2.$$

By comparing this identity with (3.47) (which is an equality as $r\geq\frac{11}{5}$), we conclude

$$\limsup_{N\to\infty}\int_Q \mathbb{S}^N:\mathbb{D}\boldsymbol{v}^N = \int_Q \mathbb{S}:\mathbb{D}\boldsymbol{v}.$$

The graph convergence lemma (Lemma A.6) then implies that $\mathbb{S}$ and $\mathbb{D}\boldsymbol{v}$ fulfill (3.46e). The proof for $r \geq \frac{11}{5}$ is thus complete.

The case $r \in \left(\frac{6}{5}, \frac{11}{5}\right)$. Step 1. Approximations and their validity. For $\epsilon > 0$, we look for $(\boldsymbol{v}^\epsilon, \mathbb{S}^\epsilon)$ such that

$$\boldsymbol{v}^\epsilon \in L^\infty\big(0,T;\mathbf{L}^2_{\boldsymbol{n},\mathrm{div}}\big) \cap L^{\frac{11}{5}}\big(0,T;\mathbf{W}^{1,\frac{11}{5}}_{0,\mathrm{div}}\big), \tag{3.57a}$$

$$\partial_t \boldsymbol{v}^\epsilon \in L^{\frac{11}{5}}\big(0,T;(\mathbf{W}^{1,\frac{11}{5}}_{0,\mathrm{div}})^*\big), \tag{3.57b}$$

$$\mathbb{S}^\epsilon \in L^{r'}(Q)^{3\times 3}_{\mathrm{sym}} \tag{3.57c}$$

satisfy

$$\begin{aligned}
&\int_Q \mathbb{S}^\epsilon : \mathbb{D}\boldsymbol{\varphi} + \epsilon \int_Q |\mathbb{D}\boldsymbol{v}^\epsilon|^{\frac{1}{5}}\, \mathbb{D}\boldsymbol{v}^\epsilon : \mathbb{D}\boldsymbol{\varphi} \\
&= \langle \boldsymbol{b}, \boldsymbol{\varphi}\rangle + \int_Q (\boldsymbol{v}^\epsilon \otimes \boldsymbol{v}^\epsilon) : \mathbb{D}\boldsymbol{\varphi} + \int_Q \boldsymbol{v}^\epsilon \cdot \frac{\partial \boldsymbol{\varphi}}{\partial t} + \int_\Omega \boldsymbol{v}_0 \cdot \boldsymbol{\varphi}(0,\cdot) \\
&\qquad \text{for all } \boldsymbol{\varphi} \in L^{\frac{11}{5}}\big(0,T;\mathbf{W}^{1,\frac{11}{5}}_{0,\mathrm{div}}\big) \text{ with } \boldsymbol{\varphi}(T,\cdot) = \mathbf{0}
\end{aligned} \tag{3.58}$$

and

$$\mathbb{S}^\epsilon = 2\nu_* \left(|\mathbb{D}\boldsymbol{v}^\epsilon| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}\boldsymbol{v}^\epsilon|)\tfrac{\mathbb{D}\boldsymbol{v}^\epsilon}{|\mathbb{D}\boldsymbol{v}^\epsilon|} \qquad \text{almost everywhere in } Q. \tag{3.59}$$

The existence of $(\boldsymbol{v}^\epsilon, \mathbb{S}^\epsilon)$ fulfilling (3.57)–(3.59) for arbitrary but fixed $\epsilon > 0$ can be proved in the same way as the existence of a weak solution to the problem for the case $r \geq \frac{11}{5}$. In addition, by taking $\boldsymbol{\varphi} := \boldsymbol{v}^\epsilon$, we have

$$\tfrac{1}{2}\|\boldsymbol{v}^\epsilon(t)\|_2^2 - \tfrac{1}{2}\|\boldsymbol{v}_0\|_2^2 + \epsilon \int_0^t \|\mathbb{D}\boldsymbol{v}^\epsilon\|^{\frac{11}{5}}_{\frac{11}{5}} + \int_0^t \int_\Omega \mathbb{S}^\epsilon : \mathbb{D}\boldsymbol{v}^\epsilon = \big\langle \boldsymbol{b}, \boldsymbol{v}^\epsilon \chi_{(0,t)\times\Omega}\big\rangle \tag{3.60}$$

for almost all $t \in (0,T)$ (and for $t = T$) where $\mathbb{D}\boldsymbol{v}^\epsilon$ and $\mathbb{S}^\epsilon$ satisfy (3.59).

Step 2. Estimates uniform with respect to ϵ and their consequences. Since $\mathbb{S} : \mathbb{D} \geq \alpha\big(|\mathbb{D}|^r + |\mathbb{S}|^{r'}\big) - \beta$ for all $(\mathbb{S}, \mathbb{D})$ fulfilling (3.59), i.e., $(\mathbb{S}, \mathbb{D}) \in \mathcal{G}$, the energy identity (3.60) implies that

$$\{\boldsymbol{v}^\epsilon; \epsilon > 0\} \text{ is bounded in } L^\infty\big(0,T;L^2(\Omega)^3\big), \tag{3.61a}$$

$$\{\boldsymbol{v}^\epsilon; \epsilon > 0\} \text{ is bounded in } L^r\big(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}}\big), \tag{3.61b}$$

$$\{\mathbb{D}\boldsymbol{v}^\epsilon; \epsilon > 0\} \text{ is bounded in } L^r\big(0,T;L^r(\Omega)^{3\times 3}_{\mathrm{sym}}\big), \tag{3.61c}$$

$$\{\epsilon^{\frac{5}{11}}\mathbb{D}\boldsymbol{v}^\epsilon; \epsilon > 0\} \text{ is bounded in } L^{\frac{11}{5}}\big(0,T;L^{\frac{11}{5}}(\Omega)^3\big), \tag{3.61d}$$

$$\{\mathbb{S}^\epsilon; \epsilon > 0\} \text{ is bounded in } L^{r'}\big(0,T;L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}\big). \tag{3.61e}$$

Using the fact that

$$\int_0^T \langle \partial_t \boldsymbol{v}^\epsilon, \boldsymbol{\varphi}\rangle = -\int_Q \boldsymbol{v}^\epsilon \cdot \partial_t \boldsymbol{\varphi} - \int_\Omega \boldsymbol{v}_0 \cdot \boldsymbol{\varphi}(0,\cdot)$$

for all $\boldsymbol{\varphi} \in L^{\frac{11}{5}}(0,T;\mathbf{W}^{1,\frac{11}{5}}_{0,\mathrm{div}})$ with $\boldsymbol{\varphi}(T,\cdot) = \mathbf{0}$ we conclude from (3.58) and (3.61) that

$$\{\partial_t \boldsymbol{v}^\epsilon; \epsilon > 0\} \text{ is bounded in } \Big(L^{\frac{5r}{5r-6}}\big(0,T;\mathbf{W}_0^{1,\frac{5r}{5r-6}}(\Omega)\big)\Big)^*.$$

These estimates together with Korn's inequality (3.2) and the Aubin-Lions lemma lead to the existence of $\boldsymbol{v}$ and $\mathbb{S}$ such that for suitable sequences $(\boldsymbol{v}^m, \mathbb{S}^m) := (\boldsymbol{v}^{\epsilon_m}, \mathbb{S}^{\epsilon_m})$ and $m \to \infty$ the following convergences hold:

$$\begin{aligned}
\boldsymbol{v}^m &\overset{*}{\rightharpoonup} \boldsymbol{v} && *\text{-weakly in } L^\infty\big(0,T;L^2(\Omega)^3\big), \\
\boldsymbol{v}^m &\rightharpoonup \boldsymbol{v} && \text{weakly in } L^r\big(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}}\big), \\
\mathbb{D}\boldsymbol{v}^m &\rightharpoonup \mathbb{D}\boldsymbol{v} && \text{weakly in } L^r\big(0,T;L^r(\Omega)^{3\times 3}_{\mathrm{sym}}\big), && (3.62) \\
\mathbb{S}^m &\rightharpoonup \mathbb{S} && \text{weakly in } L^{r'}\big(0,T;L^{r'}(\Omega)^{3\times 3}_{\mathrm{sym}}\big), && (3.63) \\
\boldsymbol{v}^m &\to \boldsymbol{v} && \text{strongly in } L^q\big(0,T;L^q(\Omega)^3\big) \text{ for all } q \in \big[1, \tfrac{5r}{3}\big).
\end{aligned}$$

Also (3.61d) implies that

$$\epsilon_m \int_Q |\mathbb{D}\boldsymbol{v}^m|^{\frac{1}{5}}\, \mathbb{D}\boldsymbol{v}^m : \mathbb{D}\boldsymbol{\varphi} \leq \epsilon_m^{\frac{5}{11}} \left(\epsilon_m \int_Q |\mathbb{D}\boldsymbol{v}^m|^{\frac{11}{5}}\right)^{\frac{6}{11}} \|\mathbb{D}\boldsymbol{\varphi}\|_{\frac{11}{5},Q} \xrightarrow{m\to\infty} 0. \tag{3.64}$$

Let us consider (3.58) with $\boldsymbol{\varphi} \in L^q(0,T;\mathbf{W}^{1,q}_{0,\mathrm{div}})$, $q = \frac{5r}{5r-6}$, $\boldsymbol{\varphi}(T,\cdot) = \mathbf{0}$ and let $m \to \infty$. Then we easily arrive to (3.46d).

Step 3. Attainment of the constitutive equation (3.46e)**.** In order to apply the graph convergence lemma (Lemma A.6) we need to proceed in a more subtle way as $\boldsymbol{w} := \boldsymbol{v}$ is not an admissible test function in (3.46d). For $\boldsymbol{u}^m := \boldsymbol{v}^m - \boldsymbol{v}$, the following identity holds

$$\begin{aligned} -\int_Q (\boldsymbol{v}^m - \boldsymbol{v})\cdot \partial_t \boldsymbol{w} + \int_Q (\mathbb{S}^m - \mathbb{S}) : \mathbb{D}\boldsymbol{w} + \epsilon_m \int_Q |\mathbb{D}\boldsymbol{v}^m|^{\frac{1}{5}}\, \mathbb{D}\boldsymbol{v}^m : \mathbb{D}\boldsymbol{\omega} \\ = \int_Q (\boldsymbol{v}^m \otimes \boldsymbol{v}^m - \boldsymbol{v}\otimes\boldsymbol{v}) : \mathbb{D}\boldsymbol{w} \\ \text{for all } \boldsymbol{w} \in \mathcal{C}^\infty([0,T];\boldsymbol{\mathcal{C}}^\infty_{0,\mathrm{div}}) \end{aligned} \tag{3.65}$$

and

$$\begin{aligned} &\boldsymbol{u}^m \overset{*}{\rightharpoonup} \mathbf{0} && *\text{-weakly in } L^\infty(0,T;\mathbf{L}^2_{\boldsymbol{n},\mathrm{div}}),\\ &\boldsymbol{u}^m \rightharpoonup \mathbf{0} && \text{weakly in } L^r(0,T;\mathbf{W}^{1,r}_{0,\mathrm{div}}),\\ &\boldsymbol{u}^m \to \mathbf{0} && \text{strongly in } L^q(0,T;L^q(\Omega)^3) \text{ for all } q \in \left[1,\tfrac{5r}{3}\right). \end{aligned}$$

We also observe that besides (3.64)

$$\begin{aligned} (\mathbb{S}^m - \mathbb{S}) =: \mathbb{H}_1^m \rightharpoonup \mathbb{O} \qquad & \text{weakly in } L^{r'}(Q),\\ \begin{pmatrix} \epsilon_m |\mathbb{D}\boldsymbol{v}^m|^{\frac{1}{5}}\, \mathbb{D}\boldsymbol{v}^m + \\ + (\boldsymbol{v}\otimes\boldsymbol{v} - \boldsymbol{v}^m \otimes \boldsymbol{v}^m) \end{pmatrix} =: \mathbb{H}_2^m \to \mathbb{O} \qquad & \text{strongly in } L^\sigma(Q) \text{ for some } \sigma \in \left(1,\tfrac{5r}{6}\right) \end{aligned}$$

and we rewrite (3.65) as

$$\int_Q \boldsymbol{u}^m \cdot \partial_t \boldsymbol{w} = \int_Q \mathbb{H}_1^m : \mathbb{D}\boldsymbol{w} + \int_Q \mathbb{H}_2^m : \mathbb{D}\boldsymbol{w}.$$

Let $Q_0 \subset Q$ be any parabolic cylinder. Take $\zeta \in \mathcal{C}_0^\infty(\frac{1}{6}Q_0)$ such that

$$\chi_{\frac{1}{8}Q_0} \leq \zeta \leq \chi_{\frac{1}{6}Q_0}.$$

Then applying the Lipschitz truncation (Lemma A.5) we conclude, using the above convergences, that

$$\limsup_{m\to\infty} \left| \int_{\frac{1}{8}Q_0 \setminus \mathcal{O}^{m,k}} (\mathbb{S}^m - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^m - \mathbb{D}\boldsymbol{v}) \right| \leq C2^{-k}.$$

This, together with the property (h) of the truncation lemma (Lemma A.5) and Hölder's inequality, implies that

$$\limsup_{m\to\infty} \int_{\frac{1}{8}Q_0} \left|(\mathbb{S}^m - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^m - \mathbb{D}\boldsymbol{v})\right|^{\frac{1}{2}} \leq C2^{-k}.$$

Set $g^m := \left|(\mathbb{S}^m - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^m - \mathbb{D}\boldsymbol{v})\right|$. Clearly $g^m \geq 0$ and $g^m \to 0$ almost everywhere in $\frac{1}{8}Q_0$. But as Q_0 is arbitrary, we conclude that

$$g^m \to 0 \text{ almost everywhere in } Q. \tag{3.66}$$

Since $\{g^m\}_{m=1}^\infty$ is bounded in $L^1(Q)$ and has pointwise limit (3.66), Corollary A.3, a consequence of the biting lemma (Lemma A.2), ensures the existence of a subsequence $\{g^{m_j}\}_{j=1}^\infty$ and a sequence of sets $\{E_k\}_{k=1}^\infty$ with $Q \supset E_1 \supset E_2 \supset \ldots$, $|E_k| \to 0$ such that for all $k \in \mathbb{N}$

$$g^{m_j} \to 0 \text{ strongly in } L^1(Q \setminus E_k).$$

From the definition of g^{m_j} we conclude that

$$\limsup_{j\to 0} \int_{Q\setminus E_k} (\mathbb{S}^{m_j} - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^{m_j} - \mathbb{D}\boldsymbol{v}) = 0,$$

which implies with the help of (3.62) and (3.63) that for all $k \in \mathbb{N}$

$$\limsup_{j\to 0} \int_{Q\setminus E_k} \mathbb{S}^{m_j} : \mathbb{D}\boldsymbol{v}^{m_j} = \int_{Q\setminus E_k} \mathbb{S} : \mathbb{D}\boldsymbol{v}.$$

Since $|E_k| \to 0$, we can conclude from the graph convergence lemma (Lemma A.6) that

$$(\mathbb{S}, \mathbb{D}\boldsymbol{v}) \in \mathcal{G} \text{ almost everywhere in } Q$$

so that (3.46e) holds.

The energy inequality and the initial condition. Since $(\mathbb{S}, \mathbb{D}\boldsymbol{v}) \in \mathcal{G}$ and $\mathcal{G}$ is monotone, we first observe that $g^m = (\mathbb{S}^m - \mathbb{S}) : (\mathbb{D}\boldsymbol{v}^m - \mathbb{D}\boldsymbol{v}) \geq 0$. It thus follows from (3.66), Fatou's lemma applied to the functions $\{g^m\}_{m=1}^{\infty}$, that are non-negative, and from the weak convergences (3.62) and (3.63) that

$$\int_Q \mathbb{S} : \mathbb{D}\boldsymbol{v} \leq \liminf_{n\to\infty} \int_Q \mathbb{S}^m : \mathbb{D}\boldsymbol{v}^m.$$

It is then easy to conclude the energy inequality (3.47) from (3.60).

Attainment of the initial condition (3.46c), which is proved with the help of the energy inequality (3.47), is standard and we omit it; see [58, sections B.3.8–10]. Thus the first part of the theorem is proved.

On the pressure. Let us consider for fixed $t \in (0, T)$ the functionals

$$\begin{aligned} \big\langle \boldsymbol{F}^1(t), \boldsymbol{\varphi} \big\rangle &:= \int_\Omega \big(\boldsymbol{v}(t,\cdot) - \boldsymbol{v}_0\big) \cdot \boldsymbol{\varphi}, \\ \big\langle \boldsymbol{F}^2(t), \boldsymbol{\varphi} \big\rangle &:= \int_\Omega \textstyle\int_0^t \big(\mathbb{S} - \boldsymbol{v} \otimes \boldsymbol{v}\big) : \mathbb{D}\boldsymbol{\varphi} - \Big\langle \textstyle\int_0^t \boldsymbol{b}, \boldsymbol{\varphi} \Big\rangle \end{aligned} \tag{3.67}$$

for $\boldsymbol{\varphi} \in \mathbf{W}_0^{1,q}$ with $q := \max\big\{r, \frac{5r}{5r-6}\big\}$. Clearly $\boldsymbol{F}^1(t), \boldsymbol{F}^2(t) \in \big(\mathbf{W}_0^{1,q}\big)^*$ for almost every $t \in (0, T)$. Testing (3.46d) by $\boldsymbol{w}^j \in \mathcal{C}_0^\infty\big([0,T); \boldsymbol{\mathcal{C}}_{0,\mathrm{div}}^\infty\big)$ such that $\boldsymbol{w}^j \to \boldsymbol{w}$ and

$$\boldsymbol{w}(s,x) = \begin{cases} \boldsymbol{\varphi}(x) & s \in [0,t), \\ \mathbf{0} & s \in [t,T) \end{cases}$$

with arbitrary $\boldsymbol{\varphi} \in \boldsymbol{\mathcal{C}}_{0,\mathrm{div}}^\infty$ and comparing with (3.67) we obtain

$$\Big\langle \big(\boldsymbol{F}^1 + \boldsymbol{F}^2\big)(t), \boldsymbol{\varphi} \Big\rangle = 0 \qquad \text{for all } \boldsymbol{\varphi} \in \boldsymbol{\mathcal{C}}_{0,\mathrm{div}}^\infty \text{ and a.e. } t \in (0,T). \tag{3.68}$$

Now consider the Stokes problems

$$-\Delta \boldsymbol{U}^1 + \nabla P^1 = \boldsymbol{F}^1, \qquad \operatorname{div} \boldsymbol{U}^1 = 0 \quad \text{in } Q, \qquad \boldsymbol{U}^1 = \mathbf{0} \quad \text{on } (0,T) \times \partial\Omega, \tag{3.69a}$$

$$-\Delta \boldsymbol{U}^2 + \nabla P^2 = \boldsymbol{F}^2, \qquad \operatorname{div} \boldsymbol{U}^2 = 0 \quad \text{in } Q, \qquad \boldsymbol{U}^2 = \mathbf{0} \quad \text{on } (0,T) \times \partial\Omega. \tag{3.69b}$$

By virtue of the assumptions on the domain, we conclude from (A.4) of Lemma A.13 that

$$\begin{aligned} \|\nabla \boldsymbol{U}^1(t)\|_6 + \|P^1(t)\|_6 &\leq C\|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_{-1,6} \leq C\|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_2, \\ \|\nabla \boldsymbol{U}^2(t)\|_{q'} + \|P^2(t)\|_{q'} &\leq C\|\boldsymbol{F}^2(t)\|_{-1,q'} \\ &\leq C\int_0^t \|\mathbb{S} - \boldsymbol{v} \otimes \boldsymbol{v}\|_{q'} + C\int_0^t \|\boldsymbol{b}\|_{-1,q'}, \end{aligned}$$

which leads to $P^1 \in L^\infty(0,T;L^6)$ and $p^2 := \partial_t P^2 \in L^{q'}(Q)$. Testing (3.69a) with $\nabla\phi$ for $\phi \in \mathcal{C}_0^\infty(\Omega)$ we get

$$\underbrace{(\nabla \boldsymbol{U}^1(t), \nabla^2\phi)}_{(\operatorname{div} \boldsymbol{U}^1(t), \Delta\phi) = 0} - (P^1(t), \Delta\phi) = (\boldsymbol{v}(t) - \boldsymbol{v}_0, \nabla\phi) = -(\operatorname{div}(\boldsymbol{v}(t) - \boldsymbol{v}_0), \phi) = 0$$

so that $(P^1(t), \Delta\phi) = 0$ for all $\phi \in \mathcal{C}_0^\infty(\Omega)$ and Weyl's lemma (cf. [81, Lemma 2], [33, Chapter 10]) yields that $P^1(t)$ is harmonic.

Testing (3.69) by $\boldsymbol{\varphi} \in \mathbf{W}^{1,q}_{0,\mathrm{div}}$ we obtain, by using (3.68),

$$(\nabla(\boldsymbol{U}^1(t) + \boldsymbol{U}^2(t)), \nabla\boldsymbol{\varphi}) = 0 \qquad \text{for all } \boldsymbol{\varphi} \in \mathbf{W}^{1,q}_{0,\mathrm{div}} \text{ and a.e. } t \in (0,T),$$

which shows together with $\boldsymbol{U}^1(t) + \boldsymbol{U}^2(t) \in \mathbf{W}^{1,q'}_{0,\mathrm{div}}$ that $\boldsymbol{U}^1 + \boldsymbol{U}^2 = \mathbf{0}$. Now we are in a position to sum up (3.69), test by $\partial_t \boldsymbol{w}$ with $\boldsymbol{w} \in \mathcal{C}_0^\infty([0,T); \boldsymbol{\mathcal{C}}^\infty_{0,\mathrm{div}})$, integrate over Q, and use the facts shown above and $P^2(0) = 0$ to obtain (3.48).

Furthermore, when Ω is a $C^{1,1}$ domain, (A.5) from Lemma A.13 yields

$$\begin{aligned} \|\nabla^2 \boldsymbol{U}^1(t)\|_2 \ &+ \|\nabla P^1(t)\|_2 \ \leq C\|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_2, \\ \|\nabla^2 \boldsymbol{U}^1(t)\|_{\frac{5r}{3}} &+ \|\nabla P^1(t)\|_{\frac{5r}{3}} \leq C\|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_{\frac{5r}{3}} \end{aligned}$$

so that $\operatorname{ess\,sup}_{t\in(0,T)} \|\nabla P^1(t)\|_2 \leq C \operatorname{ess\,sup}_{t\in(0,T)} \|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_2$ and $\int_0^T \|\nabla P^1\|_{\frac{5r}{3}}^{\frac{5r}{3}} \leq \int_0^T \|\boldsymbol{v} - \boldsymbol{v}_0\|_{\frac{5r}{3}}^{\frac{5r}{3}} \leq C$ and the proof is complete. □

3.3.2 Slip case

Here we consider the boundary condition

$$\boldsymbol{v} \cdot \boldsymbol{n} = 0 \quad \text{and} \quad \boldsymbol{s} = \mathbf{0} \qquad \text{on } (0,T) \times \partial\Omega \tag{3.70}$$

or the boundary condition

$$\boldsymbol{v} \cdot \boldsymbol{n} = 0 \quad \text{and} \quad (\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) \in \mathcal{B} \qquad \text{on } (0,T) \times \partial\Omega \tag{3.71}$$

where $\mathcal{B}$ fulfills ($\mathcal{B}$1)–($\mathcal{B}$4). The following result holds.

Theorem 3.5. *Let $T \in (0,\infty)$, $\Omega \subset \mathbb{R}^3$ be a $C^{0,1}$ domain, and $Q \coloneqq (0,T) \times \Omega$. Let $r > \frac{6}{5}$, $\boldsymbol{b} \in L^{r'}\big(0,T; (\mathbf{W}^{1,r}_{\boldsymbol{n}})^*\big)$, and $\boldsymbol{v}_0 \in \mathbf{L}^2_{\boldsymbol{n},\mathrm{div}}$. Let $\mathcal{G} \subset \mathbb{R}^{3\times3}_{\mathrm{sym}} \times \mathbb{R}^{3\times3}_{\mathrm{sym}}$ be a maximal monotone r-graph of the form* (3.10) *fulfilling ($\mathcal{G}$1)–($\mathcal{G}$4).*

(i) (Boundary condition (3.71)) *Let $\mathcal{B} \subset \mathbb{R}^3 \times \mathbb{R}^3$ be a maximal monotone 2-graph fulfilling ($\mathcal{B}$1)–($\mathcal{B}$4). Then there exists a triplet $(\boldsymbol{v}, \mathbb{S}, \boldsymbol{s})$ satisfying*

$$\boldsymbol{v} \in L^\infty(0,T; \mathbf{L}^2_{\boldsymbol{n},\mathrm{div}}) \cap L^r(0,T; \mathbf{W}^{1,r}_{\boldsymbol{n},\mathrm{div}}), \tag{3.72a}$$

$$\mathbb{S} \in L^{r'}(Q)^{3\times3}_{\mathrm{sym}}, \tag{3.72b}$$

$$\boldsymbol{s} \in L^2\big((0,T) \times \partial\Omega\big)^3, \tag{3.72c}$$

and

$$\lim_{t\to0+} \int_\Omega |\boldsymbol{v}(t,\cdot) - \boldsymbol{v}_0|^2 = 0, \tag{3.72d}$$

$$\begin{aligned} \int_Q \mathbb{S} : \mathbb{D}\boldsymbol{w} + \int\limits_{(0,T)\times\partial\Omega} \boldsymbol{s} \cdot \boldsymbol{w} &= \int_0^T \langle \boldsymbol{b}, \boldsymbol{w} \rangle + \int_Q \boldsymbol{v} \otimes \boldsymbol{v} : \mathbb{D}\boldsymbol{w} + \int_Q \boldsymbol{v} \cdot \frac{\partial \boldsymbol{w}}{\partial t} \\ &+ \int_\Omega \boldsymbol{v}_0 \cdot \boldsymbol{w}(0,\cdot) \qquad \begin{array}{c} \textit{for all } \boldsymbol{w} \in \mathcal{C}_0^\infty\big([0,T); \mathbf{W}^{1,q}_{\boldsymbol{n},\mathrm{div}}\big), \\ q = \max\big\{r, \frac{5r}{5r-6}\big\} \end{array} \end{aligned} \tag{3.72e}$$

$$\mathbb{S} = 2\nu_* \left(|\mathbb{D}\boldsymbol{v}| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}\boldsymbol{v}|) \tfrac{\mathbb{D}\boldsymbol{v}}{|\mathbb{D}\boldsymbol{v}|} \quad \textit{almost everywhere in } Q, \tag{3.72f}$$

$$(\boldsymbol{s}, \boldsymbol{w}_{\boldsymbol{\tau}}) \in \mathcal{B} \textit{ almost everywhere in } (0,T) \times \partial\Omega. \tag{3.73}$$

(ii) (Boundary condition (3.70)) *There exists a couple* $(\boldsymbol{v}, \mathbb{S})$ *satisfying* (3.72a), (3.72b), (3.72d)*,* (3.72f)*, and* (3.72e) *with* $\boldsymbol{s} = \mathbf{0}$*.*

Moreover, the following energy inequality holds:

$$\int_\Omega \frac{|\boldsymbol{v}(t,\cdot)|^2}{2} + \int_0^t \int_\Omega \mathbb{S} : \mathbb{D}\boldsymbol{v} + \int_0^t \int_{\partial\Omega} \boldsymbol{s}\cdot\boldsymbol{v} \le \int_\Omega \frac{|\boldsymbol{v}_0|^2}{2} + \int_0^t \langle \boldsymbol{b}, \boldsymbol{v}\rangle \qquad (3.74)$$
$$\text{for almost all } t \in (0,T) \text{ and for } t = T;$$

if $r \ge \frac{11}{5}$*, then* (3.74) *becomes an equality.*

In addition, if Ω *is a* $C^{1,1}$ *domain, then there is* $p \in L^{q'}(Q)$ *with* $q = \max\{r, \frac{5r}{5r-6}\}$ *such that* $\int_\Omega p(t,\cdot) = 0$ *for almost every* $t \in (0,T)$ *and*

$$\int_Q \mathbb{S} : \mathbb{D}\boldsymbol{\omega} + \int\limits_{(0,T)\times\partial\Omega} \boldsymbol{s}\cdot\boldsymbol{w} = \int_0^T \langle \boldsymbol{b}, \boldsymbol{\omega}\rangle + \int_Q \boldsymbol{v}\otimes\boldsymbol{v} : \mathbb{D}\boldsymbol{\omega} + \int_Q \boldsymbol{v}\cdot\frac{\partial\boldsymbol{\omega}}{\partial t} + \int_\Omega \boldsymbol{v}_0\cdot\boldsymbol{\omega}(0,\cdot)$$
$$+ \int_Q p \operatorname{div}\boldsymbol{\omega} \quad \text{for all } \boldsymbol{\omega} \in \mathcal{C}_0^\infty\big([0,T); \mathbf{W}^{1,q}_{\boldsymbol{n}}\big).$$

Remark 3.6. *(1) In the case of the boundary condition* (3.70) *we could define the weak solution to the problem considered differently. We could say that* $\boldsymbol{v}$ *is a weak solution to the problem if* $\boldsymbol{v}$ *fulfills* (3.72a)*,* (3.72d)*, and*

$$\int_Q 2\nu_* \left(|\mathbb{D}\boldsymbol{v}| - \delta_*\right)^+ \mathcal{S}(|\mathbb{D}\boldsymbol{v}|)\tfrac{\mathbb{D}\boldsymbol{v}}{|\mathbb{D}\boldsymbol{v}|} : \mathbb{D}\boldsymbol{w} = \int_0^T \langle \boldsymbol{b}, \boldsymbol{w}\rangle + \int_Q \boldsymbol{v}\otimes\boldsymbol{v} : \mathbb{D}\boldsymbol{w}$$
$$+ \int_Q \boldsymbol{v}\cdot\frac{\partial\boldsymbol{w}}{\partial t} + \int_\Omega \boldsymbol{v}_0\cdot\boldsymbol{w}(0,\cdot)$$
$$\text{for all } \boldsymbol{w} \in \mathcal{C}_0^\infty\big([0,T); \mathbf{W}^{1,q}_{\boldsymbol{n},\mathrm{div}}\big),\ q = \max\{r, \tfrac{5r}{5r-6}\}.$$

(2) It holds $\boldsymbol{v} \in \mathcal{C}\big(0,T; L^2(\Omega)^3\big)$ *if* $r \ge \frac{11}{5}$ *and* $\boldsymbol{v} \in \mathcal{C}\big(0,T; L^2_{\mathrm{weak}}(\Omega)^3\big)$ *if* $r \in (\frac{6}{5}, \frac{11}{5})$*.*

Proof of Theorem 3.5. We focus only on the details in which the proof differs from the proof of Theorem 3.3. Note however that a remarkable difference concerns the pressure: for the no-slip boundary condition the pressure is not integrable up to the boundary; here, for $C^{1,1}$ domains, we establish the existence of the pressure belonging to $L^s(Q)$ for some $s > 1$. This concerns in particular the no-slip/Navier-slip boundary condition which "approximates" well the no-slip boundary condition and in addition its mathematical theory admits integrable pressure.

Regarding the case $r \ge \frac{11}{5}$, the main departures from the problem with the no-slip boundary condition is due to the choice of function spaces and due to the formulation of the eigenvalue problem that generates the basis for Galerkin approximations. Here, we look for $\lambda \in \mathbb{R}$ and $\boldsymbol{\omega} \in \mathbf{W}^{3,2}_{\boldsymbol{n},\mathrm{div}} \hookrightarrow W^{1,\infty}(\Omega)^3$ satisfying

$$(\!(\boldsymbol{\omega}, \boldsymbol{\varphi})\!) = \lambda(\boldsymbol{\omega}, \boldsymbol{\varphi}) \text{ for all } \boldsymbol{\varphi} \in \mathbf{W}^{3,2}_{\boldsymbol{n},\mathrm{div}},$$

where $(\cdot,\cdot)$ is again the scalar product in $L^2(\Omega)^3$ and $(\!(\cdot,\cdot)\!)$ is a scalar product in $\mathbf{W}^{3,2}_{\boldsymbol{n},\mathrm{div}}$ defined through $(\!(\boldsymbol{\omega}, \boldsymbol{\varphi})\!) := (\nabla^3\boldsymbol{\omega}, \nabla^3\boldsymbol{\varphi}) + (\boldsymbol{\omega}, \boldsymbol{\varphi}) + (\boldsymbol{\omega}_{\boldsymbol{\tau}}, \boldsymbol{\varphi}_{\boldsymbol{\tau}})_{\partial\Omega}$. The properties of the eigenfunctions are the same as in (3.49) and consequently, for the free-slip boundary condition (3.70) there is no other change in the proof.

If the other slipping conditions are considered, then we regularize the boundary conditions as in the time independent case. Independent of the approximation parameter, we, in addition to standard uniform estimates, know that $\{\boldsymbol{s}^n\}_{n=1}^\infty$ and $\{\boldsymbol{v}^n_{\boldsymbol{\tau}}\}_{n=1}^\infty$ are bounded in $L^2(0,T; L^2(\partial\Omega)^3)$. Furthermore, as $W^{1,r}(\Omega)$ compactly embeds into $W^{\frac{1}{r},q}(\partial\Omega)$ for all $q < r$, we conclude that

$$\boldsymbol{v}^N_{\boldsymbol{\tau}} \to \boldsymbol{v}_{\boldsymbol{\tau}} \quad \text{strongly in } L^r\big(0,T; L^1(\partial\Omega)^3\big).$$

Then (up to a subsequence which we do not relabel)

$$\boldsymbol{v}^N_{\boldsymbol{\tau}} \to \boldsymbol{v}_{\boldsymbol{\tau}} \quad \text{a.e. on } (0,T)\times\partial\Omega$$

and by Egorov's theorem, for any $\delta > 0$,

$$\boldsymbol{v}_{\boldsymbol{\tau}}^N \to \boldsymbol{v}_{\boldsymbol{\tau}} \quad \text{strongly in } L^\infty(\mathcal{U}_\delta)$$

where $\mathcal{U}_\delta \subset (0,T) \times \partial\Omega$ is such that $|(0,T) \times \partial\Omega \setminus \mathcal{U}_\delta| < \delta$. The last convergence implies that

$$\limsup_{N\to\infty} \int_{\mathcal{U}_\delta} \boldsymbol{s}^N \cdot \boldsymbol{v}_{\boldsymbol{\tau}}^N = \int_{\mathcal{U}_\delta} \boldsymbol{s} \cdot \boldsymbol{v}_{\boldsymbol{\tau}}.$$

Consequently, by Lemma A.6, $(\boldsymbol{s}, \boldsymbol{v}_{\boldsymbol{\tau}}) \in \mathcal{B}$ a.e. on $\mathcal{U}_\delta$. This is true for all $r > 1$ and gives (3.73).

If $r \in (\frac{6}{5}, \frac{9}{5})$, the proof of $(\mathbb{S}, \mathbb{D}\boldsymbol{v}) \in \mathcal{G}$ is carried out as in the no-slip case, as the proof is based on local analysis in the interior of Ω.

Finally, we reconstruct the pressure. We set $p = p_1 + p_2$ where $p_1 \in L^{\frac{5r}{6}}(Q)$ solves

$$\begin{aligned} &(p_1, -\Delta z) = (\boldsymbol{v} \otimes \boldsymbol{v}, \nabla^2 z) && \text{for all } z \in W^{2,\frac{5r}{5r-6}} \text{ with } \nabla z \in \mathbf{W}_{\boldsymbol{n}}^{1,\frac{5r}{5r-6}}, \\ &\int_\Omega p_1(t) = 0 && \text{for a.e. } t \in (0,T) \end{aligned}$$

and $p_2 \in L^{r'}(Q)$ solves

$$\begin{aligned} &(p_2, -\Delta z) = \langle \boldsymbol{b}, \nabla z\rangle - (\mathbb{S}, \mathbb{D}\nabla z) - (\boldsymbol{s}, \nabla z)_{\partial\Omega} && \text{for all } z \in W^{2,\frac{5r}{5r-6}},\ \nabla z \in \mathbf{W}_{\boldsymbol{n}}^{1,\frac{5r}{5r-6}}, \\ &\int_\Omega p_2(t) = 0 && \text{for a.e. } t \in (0,T). \end{aligned}$$

Note that this is a well-posed definition because of the $C^{1,1}$ regularity of the domain Ω and Lemma A.11. Now consider a test function $\boldsymbol{\varphi} \in L^q(0,T; \mathbf{W}_{\boldsymbol{n}}^{1,q})$ and its Helmholtz decomposition using Corollary A.12:

$$\boldsymbol{\varphi} = \nabla\phi + \boldsymbol{\varphi}_0 \quad \text{with } \nabla\phi \in L^q(0,T; \mathbf{W}_{\boldsymbol{n}}^{1,q}),\ \boldsymbol{\varphi}_0 \in L^q(0,T; \mathbf{W}_{\boldsymbol{n},\mathrm{div}}^{1,q}).$$

Then we have

$$\begin{aligned} &\left\langle \frac{\partial \boldsymbol{v}}{\partial t}, \boldsymbol{\varphi} \right\rangle - (p, \operatorname{div} \boldsymbol{\varphi}) \\ &= \left\langle \frac{\partial \boldsymbol{v}}{\partial t}, \nabla\phi + \boldsymbol{\varphi}_0 \right\rangle - (p, \operatorname{div}(\nabla\phi + \boldsymbol{\varphi}_0)) = \left\langle \frac{\partial \boldsymbol{v}}{\partial t}, \boldsymbol{\varphi}_0 \right\rangle + (p_1 + p_2, -\Delta\phi) \\ &= (\boldsymbol{v} \otimes \boldsymbol{v}, \mathbb{D}\boldsymbol{\varphi}_0) - (\mathbb{S}, \mathbb{D}\boldsymbol{\varphi}_0) - \langle \boldsymbol{s}, \boldsymbol{\varphi}_0\rangle_{\partial\Omega} + \langle \boldsymbol{b}, \boldsymbol{\varphi}_0\rangle + (\boldsymbol{v} \otimes \boldsymbol{v}, \nabla^2\phi) \\ &\quad + \langle \boldsymbol{b}, \nabla\phi\rangle - (\mathbb{S}, \mathbb{D}\nabla\phi) - \langle \boldsymbol{s}, \nabla\phi\rangle_{\partial\Omega} \\ &= (\boldsymbol{v} \otimes \boldsymbol{v}, \mathbb{D}\boldsymbol{\varphi}) - (\mathbb{S}, \mathbb{D}\boldsymbol{\varphi}) - \langle \boldsymbol{s}, \boldsymbol{\varphi}_{\boldsymbol{\tau}}\rangle_{\partial\Omega} + \langle \boldsymbol{b}, \boldsymbol{\varphi}\rangle \end{aligned}$$

for a.a. $t \in (0,T)$. Thus the theorem is proven. □

4 Concluding remarks

We have classified incompressible fluids that span the gamut from Euler fluids – Navier-Stokes fluid – power-law fluids – generalized power-law fluids – stress power-law fluids – to fluids that only undergo rigid motions, that can change their constitutive character due to an activation criterion based on the value of the norm of the symmetric part of the velocity gradient or the shear stress. In the process we came across constitutive relations that have hitherto been unrecognized but could possibly be useful. In the course of our investigation we have delineated how an Euler fluid is different from a fluid that behaves like an Euler fluid prior activation and behaves like a viscous fluid when the activation criterion takes place. The latter fluid would lead to governing equations that imbed the boundary layer equations as a special case, the philosophy behind the development of the boundary layer equations and the equations governing the activated fluid being totally different. We have touched upon one important aspect in this study, namely the tremendously different properties that are exhibited by the Euler fluids and the activated Euler fluids. It is known that while the Euler fluid exhibits pathological features (such as existence of a nontrivial solution to internal flows with zero initial data and vanishing external body forces), we have shown that the new class of activated Euler fluids admits a weak solution that might be even unique in its dependence of what kind of response occurs after activation.

A classification similar to that presented here for incompressible fluids can be carried out within the context of compressible fluids, where however the framework is more complicated as there are two type of viscosities (bulk and shear) and corresponding fluidities. This issues will be addressed in a subsequent study.

Appendix A Auxiliary convergence tools

In this section, we state, without proofs, several characterizations of weak compactness in L^1. Then, following [13], we summarize several properties of refined (divergence-free) Lipschitz approximations of (divergence-free) Sobolev and Bochner-Sobolev functions. Next, we present a convergence lemma (proved recently in [21]) regarding stability of maximal monotone constitutive equations (maximal monotone r-graphs) with respect to weakly converging sequences. Finally, we close this section by the Nečas theorem and Sobolev regularity results for the Neumann-Poisson problem and the Stokes system.

In the following lemma, several assertions characterizing weak compactness in L^1, namely the Dunford-Pettis criterion (ii), uniform integrability (iii), and the de la Vallé-Poussin criterion (iv), are provided. The exact statement is taken from [29, p. 21, Theorem 10].

Lemma A.1 (Characterization of weak compactness in L^1)**.** *Let $Q \subset \mathbb{R}^M$ be a bounded measurable set and $\mathcal{V} \subset L^1(Q)$. Then the following conditions are equivalent:*

(i) any sequence $\{v_n\}_{n=1}^\infty \subset \mathcal{V}$ contains a subsequence weakly converging in $L^1(Q)$;

(ii) for any $\epsilon > 0$ there exists $K > 0$ such that for all $v \in \mathcal{V}$

$$\int_{\{|v| \geq K\}} |v(y)| \mathrm{d}y \leq \epsilon;$$

(iii) for any $\epsilon > 0$ there exists $\delta > 0$ such that for all $v \in \mathcal{V}$ and for any measurable set $M \subset Q$ such that $|M| < \delta$

$$\int_M |v(y)| \mathrm{d}y < \epsilon;$$

(iv) there exists a nonnegative function $\Phi \in \mathcal{C}([0,\infty))$ fulfilling

$$\lim_{z \to \infty} \frac{\Phi(z)}{z} = \infty,$$

such that

$$\sup_{v \in \mathcal{V}} \int_Q \Phi(|v(y)|) \mathrm{d}y < \infty.$$

Since L^1 is not reflexive, weak precompactness does not follow from boundedness. Instead bounded sequences in L^1 can exhibit local concentrations weakly converging only in the space of measures. The next lemma ensures that these concentrations are located on arbitrarily small sets and when removed (by “biting”), bounded sets are L^1-weak precompact on the complement (”unbitten” part). See original reference [15] and also [8] for a simple proof and other references.

Lemma A.2 (Biting lemma)**.** *Let $Q \subset \mathbb{R}^M$ be bounded and measurable. Let $\{v_n\}_{n=1}^\infty$ be a sequence bounded in $L^1(Q)$. Then there exist a subsequence $\{v_{n_j}\}_{j=1}^\infty \subset \{v_n\}_{n=1}^\infty$, a function $v \in L^1(Q)$, and a sequence of measurable sets $\{E_k\}_{k=1}^\infty$, $Q \supset E_1 \supset E_2 \supset \dots$, $|E_k| \to 0$ such that for all $k \in \mathbb{N}$*

$$v_{n_j} \rightharpoonup v \qquad \text{weakly in } L^1(Q \setminus E_k) \text{ as } j \to \infty.$$

In the following corollary of the preceding lemmas we establish strong convergence in L^1 up to arbitrarily small sets for a pointwise null sequence bounded in L^1.

Corollary A.3. *Let the assumptions of Lemma A.2 be fulfilled. Furthermore, assume that*

$$v_n \to 0 \quad \text{a.e. in } Q \text{ as } n \to \infty.$$

Then for the sequences $\{v_{n_j}\}_{j=1}^{\infty}$ *and* $\{E_k\}_{k=1}^{\infty}$ *from Lemma A.2 and for every* $k \in \mathbb{N}$

$$v_{n_j} \to 0 \quad \text{strongly in } L^1(Q \setminus E_k) \text{ as } j \to \infty.$$

Proof. Let $k \in \mathbb{N}$ be fixed. The sequence $\{v_{n_j}\}_{j=1}^{\infty}$ provided by the biting lemma A.2 is weakly compact in $L^1(Q \setminus E_k)$ and by the lemma A.1, (ii), $\{v_{n_j}\}_{j=1}^{\infty}$ is uniformly continuous with respect to the Lebesgue measure on $Q \setminus E_k$. By the Vitali convergence theorem, the assertions follows. □

Lipschitz approximations of solenoidal Bochner-Sobolev functions is another useful tool needed in the analysis of isochoric flows. There are several variants: Acerbi and Fusco survey the basic properties of Lipschitz approximations of Sobolev functions in [1]; further refinements have been put into place, see [30, 25]. The extension to evolutionary problems goes back to [41, 42, 27]. Further extensions have been established in [17, 13].

We first state the version [13, Theorem 4.2], which is suitable for analysis of steady problems.

Lemma A.4 (Divergence-free Lipschitz truncation of Sobolev functions)**.** *Let* $B \subset \mathbb{R}^3$ *be an arbitrary ball. Let* $r \in (1,\infty)$*. Let* $\{\boldsymbol{u}^m\}_{m=1}^{\infty} \subset \mathbf{W}^{1,r}_{0,\mathrm{div}}(B)$ *be weakly converging to zero in* $\mathbf{W}^{1,r}_{0,\mathrm{div}}(B)$*.*

Then there is a double sequence $\{\lambda_{m,k}\}_{m,k=1}^{\infty} \subset (0,\infty)$ *with*

(a) $2^{2^k} \le \lambda_{m,k} \le 2^{2^{k+1}-1}$,

a double sequence of functions $\{\boldsymbol{u}^{m,k}\}_{m,k=1}^{\infty}$*, a double sequence* $\{\mathcal{O}^{m,k}\}_{m,k=1}^{\infty}$ *of measurable subsets of* $2B$*, a constant* $C > 0$*, and* $k_0 \in \mathbb{N}$ *such that for all* $k \ge k_0$ *it holds:*

(b) $\boldsymbol{u}^{m,k} \in \mathbf{W}^{1,\infty}_{0,\mathrm{div}}(2B)$ *and* $\boldsymbol{u}^{m,k} = \boldsymbol{u}^m$ *in* $2B \setminus \mathcal{O}^{m,k}$ *for all* $m \in \mathbb{N}$,

(c) $\|\nabla \boldsymbol{u}^{m,k}\|_{L^\infty(2B)} \le C\lambda_{m,k}$ *for all* $m \in \mathbb{N}$,

(d) $\boldsymbol{u}^{m,k} \to 0$ *strongly in* $L^\infty(2B)$ *as* $m \to \infty$,

(e) $\nabla \boldsymbol{u}^{m,k} \overset{*}{\rightharpoonup} 0$ *weakly-$*$ in* $L^\infty(2B)$ *as* $m \to \infty$,

(f) $(\lambda_{m,k})^r|\mathcal{O}^{m,k}| \le C2^{-k}\|\nabla \boldsymbol{u}^m\|_r^r$ *for all* $m \in \mathbb{N}$.

Next we will formulate the assertion suitable for analysis of time-dependent problems; the presented version is taken from [13].

Lemma A.5 (Divergence-free Lipschitz truncation of Bochner-Sobolev functions)**.** *Let* $Q_0 = I_0 \times B_0 \subset \mathbb{R} \times \mathbb{R}^3$ *be a space-time cylinder. Let* $1 < r < \infty$ *with* $r, r' > \sigma > 1$, $\frac{1}{r} + \frac{1}{r'} = 1$*. Assume that there are sequences of functions* $\{\boldsymbol{u}^m\}_{m=1}^{\infty}$ *and* $\{\mathbb{H}^m\}_{m=1}^{\infty}$ *such that*

$$\begin{aligned}
\operatorname{div} \boldsymbol{u}^m &= 0 \quad \text{a.e. in } Q_0,\\
\frac{\partial \boldsymbol{u}^m}{\partial t} &= -\operatorname{div} \mathbb{H}^m \quad \text{in the sense of distributions } \left(\boldsymbol{\mathcal{C}}^{\infty}_{0,\mathrm{div}}(Q_0)\right)^*,\\
\boldsymbol{u}^m &\rightharpoonup \mathbf{0} \text{ weakly in } L^r(I_0, W^{1,r}(B_0)),\\
\boldsymbol{u}^m &\to \mathbf{0} \text{ strongly in } L^\sigma(Q_0),
\end{aligned}$$

and $\mathbb{H}^m = \mathbb{H}^m_1 + \mathbb{H}^m_2$ *satisfies*

$$\begin{aligned}
\mathbb{H}^m_1 &\rightharpoonup \mathbb{O} \text{ weakly in } L^{r'}(Q_0)),\\
\mathbb{H}^m_2 &\to \mathbb{O} \text{ strongly in } L^\sigma(Q_0).
\end{aligned}$$

Then there is a double sequence $\{\lambda^{m,k}\}_{m,k=1}^{\infty} \subset (0,\infty)$ *with*

(a) $2^{2^k} \le \lambda_{m,k} \le 2^{2^{k+1}}$,

a double sequence of functions $\{\boldsymbol{u}^{m,k}\}_{m,k=1}^{\infty} \subset L^1(Q_0)^3$, *a double sequence* $\{\mathcal{O}^{m,k}\}_{m,k=1}^{\infty}$ *of measurable subsets of* Q_0, $C > 0$, *and* $k_0 \in \mathbb{N}$ *such that for all* $k \geq k_0$*:*

(b) $\boldsymbol{u}^{m,k} \in L^s(\frac{1}{4}I_0, \mathbf{W}^{1,s}_{0,\mathrm{div}}(\frac{1}{6}B_0))$ *for all* $s \in (1,\infty)$,

(c) $\operatorname{supp} \boldsymbol{u}^{m,k} \subset \frac{1}{6}Q_0$,

(d) $\boldsymbol{u}^{m,k} = \boldsymbol{u}^m$ *a.e. on* $\frac{1}{8}Q_0 \setminus \mathcal{O}^{m,k}$,

(e) $\|\nabla \boldsymbol{u}^{m,k}\|_{L^\infty(\frac{1}{4}Q_0)} \leq C\lambda^{m,k}$,

(f) $\boldsymbol{u}^{m,k} \to 0$ *strongly in* $L^\infty(\frac{1}{4}Q_0)^3$ *as* $m \to \infty$,

(g) $\nabla \boldsymbol{u}^{m,k} \rightharpoonup 0$ *weakly in* $L^s(\frac{1}{4}Q_0)^3$ *for all* $s \in (1,\infty)$ *as* $m \to \infty$,

(h) $\limsup_{m\to\infty} (\lambda^{m,k})^r |\mathcal{O}^{m,k}| \leq C2^{-k}$,

(i) $\limsup_{m\to\infty} \big|\int_{Q_0} \mathbb{H}^m : \nabla \boldsymbol{u}^{m,k} \,\mathrm{d}x\,\mathrm{d}t\big| \leq C \limsup_{m\to\infty} (\lambda^{m,k})^r |\mathcal{O}^{m,k}|$.

Moreover, if in addition $\{\boldsymbol{u}^m\}$ *is bounded in* $L^\infty(I_0, L^6(B_0))$ *then*

(j) $\limsup_{m\to\infty} \big|\int_{\frac{1}{8}Q_0\setminus\mathcal{O}^{m,k}} \mathbb{H}^m_1 : \nabla \boldsymbol{u}^m \,\xi \,\mathrm{d}x\,\mathrm{d}t\big| \leq C\,2^{-\frac{k}{r}}$, *where* $\xi \in \mathcal{C}^\infty_0\big(\frac{Q_0}{6}\big)$ *with* $\chi_{\frac{1}{8}Q_0} \leq \xi \leq \chi_{\frac{1}{6}Q_0}$.

Another tool that we use is a simple lemma concerning the stability of the constitutive equations (represented as maximal monotone r-graphs) with respect to weakly converging sequences; see [22] for a short proof.

Lemma A.6 (Graph convergence lemma)**.** *Let* $D \subset \mathbb{R}^M$ *be an arbitrary measurable set and let a graph* $\mathcal{G}$ *fulfill the assumptions (*$\mathcal{G}2$*) and (*$\mathcal{G}3$*) on page 18. Assume that, for some* $r \in (1,\infty)$,

$$\begin{aligned} (\mathbb{S}^n, \mathbb{D}^n) &\in \mathcal{G} \quad \textit{almost everywhere in } D, \\ \mathbb{D}^n &\rightharpoonup \mathbb{D} \quad \textit{weakly in } L^r(D)^{d\times d}, \\ \mathbb{S}^n &\rightharpoonup \mathbb{S} \quad \textit{weakly in } L^{\frac{r}{r-1}}(D)^{d\times d}, \\ \limsup_{n\to\infty} \int_D \mathbb{S}^n : \mathbb{D}^n &\leq \int_D \mathbb{S} : \mathbb{D}. \end{aligned}$$

Then

$$(\mathbb{S}, \mathbb{D}) \in \mathcal{G} \textit{ almost everywhere in } D.$$

Next, we state a theorem due to Nečas [65]; the following version is from [6, Corollary 2.5 ii)].

Lemma A.7 (Nečas theorem)**.** *Let* $\Omega \subset \mathbb{R}^M$ *be a domain of class* $C^{0,1}$. *Let* $r \in (1,\infty)$. *Then there exists* $\beta > 0$ *such that*

$$\|\nabla q\|_{\left(\mathbf{W}^{1,r}_0\right)^*} := \sup_{\boldsymbol{\varphi}\in\mathbf{W}^{1,r}_0} \frac{(q, \operatorname{div}\boldsymbol{\varphi})}{\|\nabla\boldsymbol{\varphi}\|_r} \geq \beta\|q\|_{r'} \quad \textit{for all } q \in L^{r'}(\Omega) \textit{ with } \int_\Omega q = 0. \tag{A.1}$$

Remark A.8. *Lemma A.7 is closely related to the results known as the Lions lemma (coined in [55]), the Babuška-Aziz inequality, or the Ladyzhenskaya-Babuška-Brezzi condition (see [7, 14]). If we set* $L^p_0 := \big\{q \in L^p(\Omega),\ \int_\Omega q = 0\big\}$, *the statement of Lemma A.7 can also be rephrased as the following:*

(i) the gradient operator $\nabla : L^{r'}_0 \to (\mathbf{W}^{1,r}_0)^*$ *is injective with closed range,*

(ii) the divergence operator $\operatorname{div} : \mathbf{W}^{1,r}_0 \to L^r_0$ *is surjective and has a continuous right inverse, i.e., there is a bounded linear operator* $\mathcal{B} : L^r_0 \to \mathbf{W}^{1,r}_0$ *such that*

$$\operatorname{div}\mathcal{B} \textit{ is identity on } L^r_0.$$

The operator $\mathcal{B}$ *is usually called the Bogovskiĭ operator due to the explicit construction by Bogovskiĭ [10, 11]. We refer the reader to [6], where these relations are discussed in detail.*

Remark A.9. *It is shown in [6, Proposition 2.10 ii)] that for the validity of the estimate of Lemma A.7 it is sufficient to assume a priori that* $q \in \big(\mathcal{C}_0^\infty(\Omega)\big)^*$, $\int_\Omega q = 0$, *and* $\nabla q \in \big(\mathbf{W}_0^{1,r}\big)^*$. *Then, provided that* Ω *is Lipschitz,* $q \in L^{r'}(\Omega)$ *and the estimate* (A.1) *holds.*

Remark A.10. *The Lipschitz condition on* Ω *in Lemma A.7 can be weakened, see for example [26].*

Now we mention few regularity results for the Neumann problem and the Stokes system.

Lemma A.11 ($W^{2,q}$-regularity of Neumann-Poisson problem)**.** *Let* Ω *be a domain of class* $C^{1,1}$. *Let* $1 < q < \infty$ *be given. Then there exists* $C > 0$ *such that for every* $f \in L^q(\Omega)$ *with* $\int_\Omega f = 0$ *there is a weak solution* $u \in W^{1,q}(\Omega)$ *of the problem*

$$-\Delta u = f \qquad \text{in } \Omega, \tag{A.2a}$$

$$\frac{\partial u}{\partial \boldsymbol{n}} = 0 \qquad \text{on } \partial\Omega, \tag{A.2b}$$

$$\int_\Omega u = 0 \tag{A.2c}$$

fulfilling $u \in W^{2,q}(\Omega)$, $\nabla u \in \mathbf{W}_{\boldsymbol{n}}^{1,q}$, *and*

$$\|\nabla^2 u\|_q \le C\|f\|_q.$$

Proof of Lemma A.11 is outlined in [5, Remark 3.2] and invokes [3, 37, 52]. As a consequence of the lemma, we get the following result concerning the Helmholtz decomposition for functions from $\mathbf{W}_{\boldsymbol{n}}^{1,q}$.

Corollary A.12 (Helmholtz decomposition)**.** *Let* Ω *be a domain of class* $C^{1,1}$. *Let* $1 < q < \infty$ *be given. Then there exists* $C > 0$ *such that the following holds. For every* $\boldsymbol{\varphi} \in \mathbf{W}_{\boldsymbol{n}}^{1,q}$ *there exists a couple* $(\phi, \boldsymbol{\varphi}_0)$ *fulfilling*

$$\phi \in W^{2,q}(\Omega), \qquad \nabla\phi \in \mathbf{W}_{\boldsymbol{n}}^{1,q}, \qquad \boldsymbol{\varphi}_0 \in \mathbf{W}_{\boldsymbol{n},\mathrm{div}}^{1,q},$$
$$\boldsymbol{\varphi} = \nabla\phi + \boldsymbol{\varphi}_0, \qquad \|\nabla^2\phi\|_q + \|\nabla\boldsymbol{\varphi}_0\|_q \le C\|\nabla\boldsymbol{\varphi}\|_q.$$

The following lemma contains certain regularity results for the Stokes system, see [32, Theorem 2.1], [6, Proposition 4.3].

Lemma A.13 (Regularity of the Stokes system)**.** *Let* $\Omega \subset \mathbb{R}^M$ *be a domain and* $1 < q < \infty$ *be given.*

If Ω *is of class* $C^{0,1}$ *with sufficiently small Lipschitz constant* $L > 0$ *(i.e.,* $L \le L_0$ *with* $L_0 > 0$ *depending only on* M *and* q*) or* Ω *is of class* C^1 *then there exists* $C_0 > 0$ *(depending on* Ω, M, q*) such that for every* $\boldsymbol{b} \in \big(\mathbf{W}_0^{1,q'}\big)^*$ *there is a unique weak solution* $(\boldsymbol{v}, p) \in \mathbf{W}_0^{1,q} \times L^q(\Omega)$ *of the problem*

$$-\Delta\boldsymbol{v} + \nabla p = \boldsymbol{b} \qquad \text{in } \Omega, \tag{A.3a}$$

$$\operatorname{div}\boldsymbol{v} = 0 \qquad \text{in } \Omega, \tag{A.3b}$$

$$\boldsymbol{v} = \mathbf{0} \qquad \text{on } \partial\Omega \tag{A.3c}$$

and the following estimate holds true

$$\|\nabla\boldsymbol{v}\|_q + \|p\|_q \le C_0\|\boldsymbol{b}\|_{\left(\mathbf{W}_0^{1,q'}\right)^*}. \tag{A.4}$$

Furthermore, if Ω *is of class* $C^{1,1}$ *then there exists* $C_1 > 0$ *(depending on* Ω, M, q*) such that for every* $\boldsymbol{b} \in L^q(\Omega)^3$ *the unique weak solution* $(\boldsymbol{v}, p) \in \mathbf{W}_0^{1,q} \times L^q(\Omega)$ *of the problem* (A.3) *fulfills additionally* $\boldsymbol{v} \in \mathbf{W}^{2,q}$, $p \in W^{1,q}(\Omega)$, *and admits the estimate*

$$\|\nabla^2\boldsymbol{v}\|_q + \|\nabla p\|_q \le C_1\|\boldsymbol{b}\|_q. \tag{A.5}$$

Proof. The first part of the lemma is exactly the statement [32, Theorem 2.1]. This statement guarantees existence of unique $(\boldsymbol{v}, p) \in \mathbf{W}_0^{1,q} \times L^q(\Omega)$ and gives (A.4) under the aforementioned conditions.

The second part, i.e., the inclusions $\boldsymbol{v} \in \mathbf{W}^{2,q}$, $p \in W^{1,q}(\Omega)$ and the estimate (A.5) follow from [6, Proposition 4.3]. Remark 4.4 therein warns that Proposition 4.3 ibid. is not an existence result and that (A.5) holds only if a solution in the appropriate spaces exists. But we know this is the case due to the first part, in virtue of [32]. □

Appendix B Examples of maximal monotone graphs

Let us consider a graph $\mathcal{G} \subset \mathbb{R}^{3\times 3}_{\mathrm{sym}} \times \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ characterized by the relationship

$$(\mathbb{S}, \mathbb{D}) \in \mathcal{G} \Leftrightarrow \mathbb{S} = \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|} \mathcal{S}(|\mathbb{D}|)\mathbb{D} \tag{B.1}$$

with

$$\text{either} \qquad \mathcal{S}(d) = (1 + d^2)^{\frac{r-2}{2}}, \tag{B.2a}$$

$$\text{or} \qquad \mathcal{S}(d) = 1 + d^{r-2}. \tag{B.2b}$$

We will prove the following statement.

Lemma B.1. *The graph $\mathcal{G}$ characterized by* (B.1) *and* (B.2a) *with some $\delta_* \geq 0$ and $r \in (1, \infty)$ is a maximal monotone r-graph fulfilling ($\mathcal{G}$1)–($\mathcal{G}$4).*

The graph $\mathcal{G}$ characterized by (B.1) *and* (B.2b) *with some $\delta_* \geq 0$ and $r \in (1, \infty)$ is a maximal monotone q-graph fulfilling ($\mathcal{G}$1)–($\mathcal{G}$4) with $q = \max\{r, 2\}$.*

Proof. (i). Clearly $(\mathbb{O}, \mathbb{O}) \in \mathcal{G}$.
(ii). Let $\mathbb{S} = \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|} \left(1 + |\mathbb{D}|^2\right)^{\frac{r-2}{2}} \mathbb{D}$ and $\mathbb{D}_s := \mathbb{D}_2 + s(\mathbb{D}_1 - \mathbb{D}_2)$ for any $\mathbb{D}_1, \mathbb{D}_2 \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$. Then

$$\begin{aligned}
&(\mathbb{S}(\mathbb{D}_1) - \mathbb{S}(\mathbb{D}_2)) : (\mathbb{D}_1 - \mathbb{D}_2) \\
&\quad = (\mathbb{D}_1 - \mathbb{D}_2) : \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} \left[\frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-2}{2}} \mathbb{D}_s \right] \mathrm{d}s \\
&\quad = |\mathbb{D}_1 - \mathbb{D}_2|^2 \int_0^1 \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-2}{2}} \mathrm{d}s \\
&\quad + \int_0^1 (\mathbb{D}_s : (\mathbb{D}_1 - \mathbb{D}_2))^2 \Bigg\{ H\left(|\mathbb{D}_s| - \delta_*\right) \frac{\left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-2}{2}}}{|\mathbb{D}_s|^2} \\
&\qquad + (r-2) \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-4}{2}} - \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|^3} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-2}{2}} \Bigg\}.
\end{aligned}$$

Since $H\left(|\mathbb{D}_s| - \delta_*\right) - \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} = \frac{\delta_*}{|\mathbb{D}_s|}$ if $|\mathbb{D}_s| > \delta_*$ (otherwise it is zero), we observe that for $r \geq 2$

$$(\mathbb{S}(\mathbb{D}_1) - \mathbb{S}(\mathbb{D}_2)) : (\mathbb{D}_1 - \mathbb{D}_2) \geq 0.$$

If $r \in (1, 2)$, then the property ($\mathcal{G}$2) follows as well from the fact

$$\begin{aligned}
&|\mathbb{D}_1 - \mathbb{D}_2|^2 \int_0^1 \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-2}{2}} \mathrm{d}s \\
&- (2-r) \int_0^1 (\mathbb{D}_s : (\mathbb{D}_1 - \mathbb{D}_2))^2 \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-4}{2}} \mathrm{d}s \\
&\qquad \geq |\mathbb{D}_1 - \mathbb{D}_2|^2 \int_0^1 \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-4}{2}} \left(1 + |\mathbb{D}_s|^2 - (2-r)|\mathbb{D}_s|^2\right) \mathrm{d}s \\
&\qquad \geq (r-1)|\mathbb{D}_1 - \mathbb{D}_2|^2 \int_0^1 \frac{(|\mathbb{D}_s| - \delta_*)^+}{|\mathbb{D}_s|} \left(1 + |\mathbb{D}_s|^2\right)^{\frac{r-2}{2}} \mathrm{d}s \geq 0.
\end{aligned}$$

This also implies the monotone property for the graph with $\mathcal{S}(d) = d^{r-2}$ and $r > 1$. Consequently, the same is true for $\mathcal{S}(d) = 1 + d^{r-2}$.

(iii). In order to show that the graph is a maximal monotone graph, we note that the assumption: $(\mathbb{S},\mathbb{D}) \in \mathbb{R}^{3\times 3}_{\mathrm{sym}} \times \mathbb{R}^{3\times 3}_{\mathrm{sym}}$

$$\big(\mathbb{S} - \tilde{\mathbb{S}}, \mathbb{D} - \tilde{\mathbb{D}}\big) \geq 0 \quad \text{for all } (\tilde{\mathbb{S}}, \tilde{\mathbb{D}}) \in \mathcal{G}$$

implies, using (B.1) and (B.2), that

$$\left(\mathbb{S} - \frac{\big(|\tilde{\mathbb{D}}| - \delta_*\big)^+}{|\tilde{\mathbb{D}}|} \mathcal{S}(|\tilde{\mathbb{D}}|)\tilde{\mathbb{D}}\right) : \big(\mathbb{D} - \tilde{\mathbb{D}}\big) \geq 0 \quad \text{for all } \tilde{\mathbb{D}} \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}. \tag{B.3}$$

Taking $\tilde{\mathbb{D}} = \mathbb{D} \pm \lambda \mathbb{A}$, $\mathbb{A}$ arbitrary, $\lambda > 0$, we conclude from (B.3) that

$$\mp \mathbb{A} : \left(\mathbb{S} - \frac{(|\mathbb{D} \pm \lambda\mathbb{A}| - \delta_*)^+}{|\mathbb{D} \pm \lambda\mathbb{A}|} \mathcal{S}(|\mathbb{D} \pm \lambda\mathbb{A}|)(\mathbb{D} \pm \lambda\mathbb{A})\right) \geq 0.$$

Letting $\lambda \to 0+$, we finally obtain (using continuity of the involved functions)

$$\left(\mathbb{S} - \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|} \mathcal{S}(|\mathbb{D}|)\mathbb{D}\right) : \mathbb{A} = 0 \quad \text{for all } \mathbb{A} \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}.$$

Hence $\mathbb{S}$ and $\mathbb{D}$ fulfill the right-hand side of (B.1) and thus $(\mathbb{S},\mathbb{D}) \in \mathcal{G}$.
(iv). Assume that $\delta_* > 0$. For $d \geq 0$ we have

$$\begin{aligned}
\min\Big\{1, \tfrac{\big(1+(2\delta_*)^2\big)^{\frac{r-2}{2}}}{(2\delta_*)^{r-2}}\Big\} H(d - 2\delta_*) d^{r-1} &\leq (1 + d^2)^{\frac{r-2}{2}} d \leq (1 + d^2)^{\frac{r-1}{2}} \\
&\leq (1 + d)^{r-1} \leq \Big((2\delta_*)^{-1} \max\{d, 2\delta_*\} + \max\{d, 2\delta_*\}\Big)^{r-1}, \\
d^{q-1} \leq (1 + d^{r-2})d &= d + d^{r-1} \\
&\leq (2\delta_*)^{2-q}\big(\max\{d, 2\delta_*\}\big)^{q-1} + (2\delta_*)^{r-q}\big(\max\{d, 2\delta_*\}\big)^{q-1}
\end{aligned} \tag{B.4}$$

where $q = \max\{r, 2\}$, $H(t) = 1$ for $t > 0$, $H(t) = 0$ otherwise. Let us define

$$q := \begin{cases} r & \text{case (B.2a)}, \\ \max\{r, 2\} & \text{case (B.2b)}. \end{cases}$$

Due to (B.4) we have, for the both cases in (B.2),

$$C_1(\delta_*, r) H(|\mathbb{D}| - 2\delta_*)|\mathbb{D}|^{q-1} \leq \mathcal{S}(|\mathbb{D}|)|\mathbb{D}| \leq C_2(\delta_*, r)\big(\max\{|\mathbb{D}|, 2\delta_*\}\big)^{q-1} \tag{B.5}$$

with certain $C_1(\delta_*, r)$, $C_2(\delta_*, r) > 0$ independent of $\mathbb{D}$. Notice also that

$$\tfrac{1}{2} H(|\mathbb{D}| - 2\delta_*) \leq \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|} \leq 1. \tag{B.6}$$

Now define $\mathbb{S} = \frac{(|\mathbb{D}|-\delta_*)^+}{|\mathbb{D}|}\mathcal{S}(|\mathbb{D}|)\mathbb{D}$ and observe that with the help of the right-wing inequalities of (B.5) and (B.6) we obtain

$$\begin{aligned}
|\mathbb{D}|^q + |\mathbb{S}|^{q'} \leq |\mathbb{D}|^q + \Big(C_2(\delta_*, r)\big(\max\{|\mathbb{D}|, 2\delta_*\}\big)^{q-1}\Big)^{q'}& \\
\leq C_3(\delta_*, r)\big(\max\{|\mathbb{D}|, 2\delta_*\}\big)^q&
\end{aligned}$$

with certain $C_3(\delta_*, r) > 0$ independent of $\mathbb{D}$ and $\mathbb{S}$. Hence, with the help of the left-wing inequalities of (B.5) and (B.6),

$$\begin{aligned}
C_3(\delta_*, r)^{-1}\big(|\mathbb{D}|^q + |\mathbb{S}|^{q'}\big) - (2\delta_*)^q &\leq \big(\max\{|\mathbb{D}|, 2\delta_*\}\big)^q - (2\delta_*)^q \\
\leq H(|\mathbb{D}| - 2\delta_*)|\mathbb{D}|^q &\leq 2C_1(\delta_*, r)^{-1} \frac{(|\mathbb{D}| - \delta_*)^+}{|\mathbb{D}|} \mathcal{S}(|\mathbb{D}|)|\mathbb{D}|^2 = 2C_1(\delta_*, r)^{-1} \mathbb{S} : \mathbb{D}
\end{aligned}$$

which is the last property $(\mathcal{G}4)$. We leave the case $\delta_* = 0$ as an exercise. □

References

[1] E. Acerbi and N. Fusco. “An approximation lemma for $W^{1,p}$ functions”. In: *Material instabilities in continuum mechanics (Edinburgh, 1985–1986).* Oxford Sci. Publ. New York: Oxford Univ. Press, 1988, pp. 1–5.

[2] R. A. Adams and J. J. F. Fournier. *Sobolev spaces.* Second. Vol. 140. Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, 2003, pp. xiv+305.

[3] S. Agmon, A. Douglis, and L. Nirenberg. “Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I”. In: *Comm. Pure Appl. Math.* 12 (1959), pp. 623–727. DOI: 10.1002/cpa.3160120405.

[4] R. Alexandre, Y.-G. Wang, C.-J. Xu, and T. Yang. “Well-posedness of the Prandtl equation in Sobolev spaces”. In: *J. Amer. Math. Soc.* 28.3 (2015), pp. 745–784. DOI: 10.1090/S0894-0347-2014-00813-4.

[5] C. Amrouche, V. Girault, and J. Giroire. “Dirichlet and Neumann exterior problems for the n-dimensional Laplace operator: an approach in weighted Sobolev spaces”. In: *J. Math. Pures Appl. (9)* 76.1 (1997), pp. 55–81. DOI: 10.1016/S0021-7824(97)89945-X.

[6] C. Amrouche and V. Girault. “Decomposition of vector spaces and application to the Stokes problem in arbitrary dimension”. In: *Czechoslovak Math. J.* 44(119).1 (1994), pp. 109–140. URL: http://hdl.handle.net/10338.dmlcz/128452.

[7] I. Babuška. “The finite element method with Lagrangian multipliers”. In: *Numer. Math.* 20 (1973), pp. 179–192. DOI: 10.1007/BF01436561.

[8] J. M. Ball and F. Murat. “Remarks on Chacon’s biting lemma”. In: *Proceedings of the American Mathematical Society* 107.3 (1989), pp. 655–663. DOI: 10.2307/2048162.

[9] S. Bauer and D. Pauly. “On Korn’s first inequality for mixed tangential and normal boundary conditions on bounded Lipschitz domains in $\mathbb{R}^N$”. In: *Ann. Univ. Ferrara Sez. VII Sci. Mat.* 62.2 (2016), pp. 173–188. DOI: 10.1007/s11565-016-0247-x.

[10] M. E. Bogovskiĭ. “Solution of the first boundary value problem for an equation of continuity of an incompressible medium”. In: *Dokl. Akad. Nauk SSSR* 248.5 (1979), pp. 1037–1040. URL: https://ci.nii.ac.jp/naid/10024332196/en/.

[11] M. E. Bogovskiĭ. “Solutions of some problems of vector analysis, associated with the operators div and grad”. In: *Theory of cubature formulas and the application of functional analysis to problems of mathematical physics.* Vol. 1980. Trudy Sem. S. L. Soboleva, No. 1. Akad. Nauk SSSR Sibirsk. Otdel., Inst. Mat., Novosibirsk, 1980, pp. 5–40, 149. URL: https://ci.nii.ac.jp/naid/10006414459/en/.

[12] P. Boltenhagen, Y. Hu, E. Matthys, and D. Pine. “Observation of bulk phase separation and coexistence in a sheared micellar solution”. In: *Physical Review Letters* 79.12 (1997), pp. 2359–2362. DOI: 10.1103/PhysRevLett.79.2359.

[13] D. Breit, L. Diening, and S. Schwarzacher. “Solenoidal Lipschitz truncation for parabolic PDEs”. In: *Mathematical Models and Methods in Applied Sciences* 23.14 (2013), pp. 2671–2700. DOI: 10.1142/S0218202513500437.

[14] F. Brezzi. “On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers”. In: *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge* 8.R-2 (1974), pp. 129–151. DOI: 10.1051/m2an/197408R201291.

[15] J. Brooks and R. Chacon. “Continuity and compactness of measures”. In: *Advances in Mathematics* 37.1 (1980), pp. 16–26. DOI: 10.1016/0001-8708(80)90023-7.

[16] M. Bulíček, P. Gwiazda, J. Málek, K. R. Rajagopal, and A. Świerczewska-Gwiazda. “On flows of fluids described by an implicit constitutive equation characterized by a maximal monotone graph”. In: *Mathematical Aspects of Fluid Mechanics.* Vol. 402. London Math. Soc. Lecture Note Ser. Cambridge: Cambridge Univ. Press, 2012, pp. 23–51. URL: https://books.google.cz/books?id=aix-XZ9yRjEC.

[17] M. Bulíček, P. Gwiazda, J. Málek, and A. Świerczewska-Gwiazda. "On Unsteady Flows of Implicitly Constituted Incompressible Fluids". In: *SIAM J. Math. Anal.* 44.4 (2012), pp. 2756–2801. DOI: `10.1137/110830289`.

[18] M. Bulíček, P. Kaplický, and D. Pražák. "Time regularity of flows of non-Newtonian fluids with critical power-law growth". Preprint. Feb. 2018.

[19] M. Bulíček, J. Málek, and K. R. Rajagopal. "Navier's slip and evolutionary Navier-Stokes-like systems with pressure and shear-rate dependent viscosity". In: *Indiana Univ. Math. J.* 56.1 (2007), pp. 51–85. DOI: `10.1512/iumj.2007.56.2997`.

[20] M. Bulíček, F. Ettwein, P. Kaplický, and D. Pražák. "On uniqueness and time regularity of flows of power-law like non-Newtonian fluids". In: *Mathematical Methods in the Applied Sciences* 33.16 (2010), pp. 1995–2010. DOI: `10.1002/mma.1314`.

[21] M. Bulíček and J. Málek. "Internal flows of incompressible fluids subject to stick-slip boundary conditions". In: *Vietnam J. Math.* (2016), pp. 1–14. DOI: `10.1007/s10013-016-0221-z`.

[22] M. Bulíček and J. Málek. "On unsteady internal flows of Bingham fluids subject to threshold slip on the impermeable boundary". In: *Recent Developments of Mathematical Fluid Mechanics*. Ed. by H. Amann, Y. Giga, H. Kozono, H. Okamoto, and M. Yamazaki. Basel: Springer Basel, 2016, pp. 135–156. DOI: `10.1007/978-3-0348-0939-9_8`.

[23] D.L. Goodstein. *States of Matter*. Dover Publications, 1985.

[24] C. De Lellis and L. Székelyhidi Jr. "On admissibility criteria for weak solutions of the Euler equations". In: *Arch. Ration. Mech. Anal.* 195.1 (2010), pp. 225–260. DOI: `10.1007/s00205-008-0201-x`.

[25] L. Diening, J. Málek, and M. Steinhauer. "On Lipschitz truncations of Sobolev functions (with variable exponent) and their selected applications". In: *ESAIM Control Optim. Calc. Var.* 14.2 (2008), pp. 211–232. DOI: `10.1051/cocv:2007049`.

[26] L. Diening, M. Růžička, and K. Schumacher. "A decomposition technique for John domains". In: *Ann. Acad. Sci. Fenn. Math.* 35.1 (2010), pp. 87–114. DOI: `10.5186/aasfm.2010.3506`.

[27] L. Diening, M. Růžička, and J. Wolf. "Existence of weak solutions for unsteady motions of generalized Newtonian fluids". In: *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* 9.1 (2010), pp. 1–46. URL: `http://www.numdam.org/item/ASNSP_2010_5_9_1_1_0`.

[28] G. Duvant and J.-L. Lions. *Inequalities in mechanics and physics*. Berlin: Springer, 1976.

[29] E. Feireisl, T. G. Karper, and M. Pokorný. *Mathematical Theory of Compressible Viscous Fluids: Analysis and Numerics*. Advances in Mathematical Fluid Mechanics. Springer International Publishing, 2016. DOI: `10.1007/978-3-319-44835-0`.

[30] J. Frehse, J. Málek, and M. Steinhauer. "On analysis of steady flows of fluids with shear-dependent viscosity based on the Lipschitz truncation method". In: *SIAM J. Math. Anal.* 34.5 (2003), 1064–1083 (electronic). DOI: `10.1137/S0036141002410988`.

[31] J. Frehse and J. Málek. "Problems Due to the No-Slip Boundary in Incompressible Fluid Dynamics". In: *Geometric Analysis and Nonlinear Partial Differential Equations*. Ed. by S. Hildebrandt and H. Karcher. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 559–571. DOI: `10.1007/978-3-642-55627-2_29`.

[32] G. P. Galdi, C. G. Simader, and H. Sohr. "On the Stokes problem in Lipschitz domains". In: *Ann. Mat. Pura Appl. (4)* 167 (1994), pp. 147–163. DOI: `10.1007/BF01760332`.

[33] L. Gårding. *Some points of analysis and their history*. Vol. 11. University Lecture Series. American Mathematical Society, Providence, RI; Higher Education Press, Beijing, 1997, pp. viii+88. DOI: `10.1090/ulect/011`.

[34] D. Gérard-Varet, Y. Maekawa, and N. Masmoudi. "Gevrey stability of Prandtl expansions for 2-dimensional Navier-Stokes flows". In: *Duke Math. J.* 167.13 (2018), pp. 2531–2631. DOI: `10.1215/00127094-2018-0020`.

[35] D. Gérard-Varet and T. Nguyen. "Remarks on the ill-posedness of the Prandtl equation". In: *Asymptot. Anal.* 77.1-2 (2012), pp. 71–88. DOI: `10.3233/ASY-2011-1075`.

[36] D. Gérard-Varet and N. Masmoudi. “Well-posedness for the Prandtl system without analyticity or monotonicity”. In: *Ann. Sci. Éc. Norm. Supér. (4)* 48.6 (2015), pp. 1273–1325. DOI: 10.24033/asens.2270.

[37] P. Grisvard. *Elliptic problems in nonsmooth domains.* Vol. 24. Monographs and Studies in Mathematics. Boston, MA: Pitman (Advanced Publishing Program), 1985, pp. xiv+410.

[38] C. Harley, E. Momoniat, and K. Rajagopal. “Reversal of flow of a non-Newtonian fluid in an expanding channel”. In: *International Journal of Non-Linear Mechanics* 101 (2018), pp. 44–55. DOI: 10.1016/j.ijnonlinmec.2018.02.006.

[39] C. Holmes, M. Cates, M. Fuchs, and P. Sollich. “Glass transitions and shear thickening suspension rheology”. In: *Journal of Rheology* 49.1 (2005), pp. 237–269. DOI: 10.1122/1.1814114.

[40] A. Janečka, J. Málek, V. Průša, and G. Tierra. “Numerical scheme for simulation of transient flows of non-Newtonian fluids characterised by a non-monotone relation between the symmetric part of the velocity gradient and the Cauchy stress tensor”. In: *Acta Mechanica* (2019). Accepted for publication. Preprint https://arxiv.org/abs/1809.04323.

[41] J. Kinnunen and J. L. Lewis. “Higher integrability for parabolic systems of p-Laplacian type”. In: *Duke Mathematical Journal* 102.2 (2000), pp. 253–272. DOI: 10.1215/S0012-7094-00-10223-2.

[42] J. Kinnunen and J. L. Lewis. “Very weak solutions of parabolic systems of p-Laplacian type”. In: *Arkiv för Matematik* 40.1 (2002), pp. 105–132. DOI: 10.1007/BF02384505.

[43] H. Koch and V. A. Solonnikov. “L^p-estimates for a solution to the nonstationary Stokes equations”. In: *Journal of Mathematical Sciences* 106.3 (2001), pp. 3042–3072. DOI: 10.1023/A:1011375706754.

[44] H. Koch and V. A. Solonnikov. “L^q-Estimates of the First-Order Derivatives of Solutions to the Nonstationary Stokes Problem”. In: *Nonlinear Problems in Mathematical Physics and Related Topics I.* Springer, 2002, pp. 203–218. DOI: 10.1007/978-1-4615-0777-2_12.

[45] I. Kukavica, N. Masmoudi, V. Vicol, and T. K. Wong. “On the local well-posedness of the Prandtl and hydrostatic Euler equations with multiple monotonicity regions”. In: *SIAM J. Math. Anal.* 46.6 (2014), pp. 3865–3890. DOI: 10.1137/140956440.

[46] I. Kukavica and V. Vicol. “On the local existence of analytic solutions to the Prandtl boundary layer equations”. In: *Commun. Math. Sci.* 11.1 (2013), pp. 269–292. DOI: 10.4310/CMS.2013.v11.n1.a8.

[47] O. A. Ladyzhenskaya. *The mathematical theory of viscous incompressible flow.* Second English edition, revised and enlarged. Translated from the Russian by Richard A. Silverman and John Chu. Mathematics and its Applications, Vol. 2. Gordon and Breach, Science Publishers, New York-London-Paris, 1969, pp. xviii+224.

[48] O. Ladyzhenskaya. “New equations for the description of motion of viscous incompressible fluids and solvability in the large of boundary value problems for them.” Russian. In: *Tr. Mat. Inst. Steklova* 102 (1967), pp. 85–104. URL: http://mi.mathnet.ru/eng/tm2939.

[49] O. Ladyzhenskaya. “Modification of the Navier-Stokes equations for large velocity gradients.” Russian. In: *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov* 7 (1968), pp. 126–154. URL: http://mi.mathnet.ru/eng/znsl2239.

[50] C. Le Roux and K. R. Rajagopal. “Shear flows of a new class of power-law fluids”. In: *Applications of Mathematics* 58.2 (2013), pp. 153–177. DOI: 10.1007/s10492-013-0008-4.

[51] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires.* Etudes mathématiques. Dunod, 1969, pp. xx+554. URL: https://books.google.cz/books?id=1VJnngEACAAJ.

[52] J.-L. Lions and E. Magenes. “Problemi ai limiti non omogenei. V”. In: *Ann. Scuola Norm Sup. Pisa (3)* 16 (1962), pp. 1–44. URL: http://www.numdam.org/item/ASNSP_1962_3_16_1_1_0.

[53] C.-J. Liu, Y.-G. Wang, and T. Yang. “On the ill-posedness of the Prandtl equations in three-dimensional space”. In: *Arch. Ration. Mech. Anal.* 220.1 (2016), pp. 83–108. DOI: 10.1007/s00205-015-0927-1.

[54] M. C. Lombardo, M. Cannone, and M. Sammartino. “Well-posedness of the boundary layer equations”. In: *SIAM J. Math. Anal.* 35.4 (2003), pp. 987–1004. DOI: 10.1137/S0036141002412057.

[55] E. Magenes and G. Stampacchia. “I problemi al contorno per le equazioni differenziali di tipo ellittico”. In: *Ann. Scuola Norm. Sup. Pisa (3)* 12 (1958), pp. 247–358. URL: http://www.numdam.org/item/ASNSP_1958_3_12_3_247_0.

[56] J. Málek, J. Nečas, M. Rokyta, and M. Růžička. *Weak and measure-valued solutions to evolutionary PDEs.* London: Chapman & Hall, 1996, pp. xii+317.

[57] J. Málek, V. Průša, and K. R. Rajagopal. “Generalizations of the Navier-Stokes fluid from a new perspective”. In: *International Journal of Engineering Science* 48.12 (2010), pp. 1907–1924. DOI: 10.1016/j.ijengsci.2010.06.013.

[58] J. Málek and K. R. Rajagopal. “Mathematical issues concerning the Navier-Stokes equations and some of its generalizations”. In: *Evolutionary equations. Vol. II.* Handb. Differ. Equ. Elsevier/North-Holland, Amsterdam, 2005, pp. 371–459. DOI: 10.1016/S1874-5717(06)80008-3.

[59] J. Málek and K. R. Rajagopal. “Compressible generalized Newtonian fluids”. In: *Z. Angew. Math. Phys.* 61.6 (2010), pp. 1097–1110. DOI: 10.1007/s00033-010-0061-8.

[60] J. Málek, K. R. Rajagopal, and M. Růžička. “Existence and regularity of solutions and the stability of the rest state for fluids with shear dependent viscosity”. In: *Math. Models Methods Appl. Sci.* 5.6 (1995), pp. 789–812. DOI: 10.1142/S0218202595000449.

[61] D. Mansutti, G. Pontrelli, and K. R. Rajagopal. “Steady flows of non-Newtonian fluids past a porous plate with suction or injection”. In: *International Journal for Numerical Methods in Fluids* 17.11 (1993), pp. 927–941. DOI: 10.1002/fld.1650171102.

[62] D. Mansutti and K. Rajagopal. “Flow of a shear thinning fluid between intersecting planes”. In: *International journal of non-linear mechanics* 26.5 (1991), pp. 769–775. DOI: 10.1016/0020-7462(91)90027-Q.

[63] E. Maringová and J. Žabenský. “On a Navier-Stokes-Fourier-like system capturing transitions between viscous and inviscid fluid regimes and between no-slip and perfect-slip boundary conditions”. In: *Nonlinear Analysis: Real World Applications* 41 (2018), pp. 152–178. DOI: 10.1016/j.nonrwa.2017.10.008.

[64] N. Masmoudi and T. K. Wong. “Local-in-time existence and uniqueness of solutions to the Prandtl equations by energy methods”. In: *Comm. Pure Appl. Math.* 68.10 (2015), pp. 1683–1741. DOI: 10.1002/cpa.21595.

[65] J. Nečas. *Direct methods in the theory of elliptic equations.* Springer Monographs in Mathematics. Translated from the 1967 French original by Gerard Tronel and Alois Kufner, Editorial coordination and preface by Šárka Nečasová and a contribution by Christian G. Simader. Springer, Heidelberg, 2012, pp. xvi+372. DOI: 10.1007/978-3-642-10455-8.

[66] O. A. Oleĭnik. “On the system of Prandtl equations in boundary-layer theory”. In: *Dokl. Akad. Nauk SSSR* 150 (1963), pp. 28–31. URL: http://mi.mathnet.ru/eng/dan27902.

[67] O. A. Oleinik and V. N. Samokhin. *Mathematical models in boundary layer theory.* Vol. 15. Applied Mathematics and Mathematical Computation. Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. x+516.

[68] T. Perlácová and V. Průša. “Tensorial implicit constitutive relations in mechanics of incompressible non-Newtonian fluids”. In: *Journal of Non-Newtonian Fluid Mechanics* 216 (2015), pp. 13–21. DOI: 10.1016/j.jnnfm.2014.12.006.

[69] L. Prandtl. “Über Flüssigkeitsbewegung bei sehr kleiner Reibung (On the motion of fluids with very little friction)”. In: *Verhandlungen des dritten internaionalen Mathematiker-Kongresses in Heidelberg 1904.* Ed. by A. Krazer. Leipzig: Teubner, 1905, pp. 484–491. DOI: 10.1007/978-3-662-11836-8_43.

[70] K. R. Rajagopal. “Boundary layers in non-linear fluids”. In: *Trends in applications of mathematics to mechanics*. Ed. by M. M. Marques and J.F. Rodrigues. Pittman Monographs and Surveys in Pure and Applied Mathematics 77. Longman, 1995, pp. 209–218.

[71] K. R. Rajagopal. “On implicit constitutive theories”. In: *Appl. Math.* 48 (2003), pp. 279–319. DOI: 10.1023/A:1026062615145.

[72] K. R. Rajagopal. “On implicit constitutive theories for fluids”. In: *Journal of Fluid Mechanics* 550 (2006), pp. 243–249. DOI: 10.1017/S0022112005008025.

[73] K. R. Rajagopal and A. R. Srinivasa. “On the thermodynamics of fluids defined by implicit constitutive relations”. In: *Z. Angew. Math. Phys.* 59 (2008), pp. 715–729. DOI: 10.1007/s00033-007-7039-1.

[74] K. R. Rajagopal and A. Wineman. “Flow of a BKZ fluid in an Orthogonal Rheometer”. In: *Journal of Rheology* 27.5 (1983), pp. 509–516. DOI: 10.1122/1.549729.

[75] K. R. Rajagopal, A. S. Gupta, and T. Y. Na. “A note on the Falkner-Skan flows of a non-Newtonian fluid”. In: *Internat. J. Non-Linear Mech.* 18.4 (1983), pp. 313–320. DOI: 10.1016/0020-7462(83)90028-8.

[76] K. R. Rajagopal, A. S. Gupta, and A. Wineman. “On a boundary layer theory for non-Newtonian fluids”. In: *International Journal of Engineering Science* 18.6 (1980), pp. 875–883. DOI: 10.1016/0020-7225(80)90035-X.

[77] K. Rajagopal. “Remarks on the notion of "pressure"”. In: *International Journal of Non-Linear Mechanics* 71 (2015), pp. 165–172. DOI: 10.1016/j.ijnonlinmec.2014.11.031.

[78] M. Sammartino and R. E. Caflisch. “Zero viscosity limit for analytic solutions of the Navier-Stokes equation on a half-space. II. Construction of the Navier-Stokes solution”. In: *Comm. Math. Phys.* 192.2 (1998), pp. 463–491. DOI: 10.1007/s002200050305.

[79] M. Sammartino and R. E. Caflisch. “Zero viscosity limit for analytic solutions, of the Navier-Stokes equation on a half-space. I. Existence for Euler and Prandtl equations”. In: *Comm. Math. Phys.* 192.2 (1998), pp. 433–461. DOI: 10.1007/s002200050304.

[80] H. Schlichting. *Boundary Layer Theory*. McGraw-Hill, 1960.

[81] H. Weyl. “The method of orthogonal projection in potential theory”. In: *Duke Math. J.* 7 (1940), pp. 411–444. URL: http://projecteuclid.org/euclid.dmj/1077492266.

[82] E. Wiedemann. “Existence of weak solutions for the incompressible Euler equations”. In: *Annales de l’Institut Henri Poincaré C, Non Linear Analysis* 28.5 (2011), pp. 727–730. DOI: 10.1016/j.anihpc.2011.05.002.

Chapter II

Localization of the $W^{-1,q}$ norm for local a posteriori efficiency[1]

1 Introduction

The weak solution of the Dirichlet problem associated with the Laplace equation is a function u characterized by

$$u - u^{\mathrm{D}} \in W_0^{1,2}(\Omega), \tag{1.1a}$$

$$(\nabla u, \nabla v) = (f, v) \qquad \forall v \in W_0^{1,2}(\Omega). \tag{1.1b}$$

Here $\Omega \subset \mathbb{R}^d$, $d \geq 1$, $f \in L^2(\Omega)$, and $u^{\mathrm{D}} \in W^{1,2}(\Omega)$. A typical numerical approximation of u gives u_h such that $u_h - u^{\mathrm{D}} \in V_h^0 \subset W_0^{1,2}(\Omega)$; we assume for simplicity that u^{D} lies in the same discrete space $V_h \subset W_0^{1,2}(\Omega)$ as u_h, so that there is no Dirichlet datum interpolation error.

The *intrinsic distance* of u_h to u is the $W_0^{1,2}(\Omega)$*-norm error* given by $\|\nabla(u - u_h)\|$. This distance is *localizable* in the sense that it is equal to a Hilbertian sum of the $W^{1,2}(\Omega)$-seminorm errors $\|\nabla(u - u_h)\|_K$ over elements K of a partition $\mathcal{T}_h$ of Ω, i.e.,

$$\|\nabla(u - u_h)\| = \left\{ \sum_{K \in \mathcal{T}_h} \|\nabla(u - u_h)\|_K^2 \right\}^{\frac{1}{2}}. \tag{1.2}$$

It is this problem-given intrinsic distance that is the most suitable for a posteriori error control. Under appropriate conditions, namely when the orthogonality $(f, \psi_{\boldsymbol{a}}) - (\nabla u_h, \nabla \psi_{\boldsymbol{a}}) = 0$ is fulfilled for the "hat" functions $\psi_{\boldsymbol{a}}$ associated with the interior vertices $\boldsymbol{a}$ of the partition $\mathcal{T}_h$, there exist a posteriori estimators $\eta_K(u_h)$, fully and locally computable from u_h, such that

$$\|\nabla(u - u_h)\| \leq \left\{ \sum_{K \in \mathcal{T}_h} \eta_K(u_h)^2 \right\}^{\frac{1}{2}} \tag{1.3}$$

and such that

$$\eta_K(u_h) \leq C \left\{ \sum_{K' \in \mathcal{T}_K} \|\nabla(u - u_h)\|_{K'}^2 \right\}^{\frac{1}{2}}, \tag{1.4}$$

where C is a generic constant and $\mathcal{T}_K$ is some local neighborhood of the element K, see Carstensen and Funken [23], Braess *et al.* [15], Veeser and Verfürth [50], Ern and Vohralík [34], or Verfürth [53] and the references therein. Property (1.4) is called *local efficiency* and is clearly only possible thanks to (1.2), the local structure of the $W_0^{1,2}(\Omega)$-norm distance. A different equivalence result where locality plays a central role is that of Veeser [49], see also [3],

[1] This chapter is a pre-copyedited, author-produced version of an article accepted for publication in *IMA Journal of Numerical Analysis* following peer review. The version of record [*Jan Blechta, Josef Málek, and Martin Vohralík. Localization of the $W^{-1,q}$ norm for local a posteriori efficiency. IMA J. Numer. Anal., 2019. Oxford University Press.*] is available online at `https://doi.org/10.1093/imanum/drz002`.

who recently proved that the local- and global-best approximation errors in the $W_0^{1,2}(\Omega)$-norm are equivalent.

Many problems are nonlinear; the basic model that represents one example of a general class of nonlinear models considered here is the Dirichlet problem associated with the p-Laplace equation, where, in place of (1.1), one looks for function u such that

$$\begin{aligned} u - u^{\mathrm{D}} &\in W_0^{1,p}(\Omega), \\ (\boldsymbol{\sigma}(\nabla u), \nabla v) &= (f, v) \qquad \forall v \in W_0^{1,p}(\Omega), \\ \boldsymbol{\sigma}(\boldsymbol{g}) &= |\boldsymbol{g}|^{p-2}\boldsymbol{g} \qquad \boldsymbol{g} \in \mathbb{R}^d \end{aligned}$$

for some $p \in (1, \infty)$, $u^{\mathrm{D}} \in W^{1,p}(\Omega)$, and $f \in L^q(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Let $u_h \in V_h \subset W^{1,p}(\Omega)$ fulfilling $u_h - u^{\mathrm{D}} \in V_h^0 \subset W_0^{1,p}(\Omega)$ be a numerical approximation of the exact solution u and let $\mathcal{R}(u_h)$ be the *residual* of u_h given by

$$\langle \mathcal{R}(u_h), v\rangle_{W^{-1,q}(\Omega), W_0^{1,p}(\Omega)} := (f, v) - (\boldsymbol{\sigma}(\nabla u_h), \nabla v) \qquad v \in W_0^{1,p}(\Omega); \tag{1.5}$$

$\mathcal{R}(u_h)$ belongs to $W^{-1,q}(\Omega) := \big(W_0^{1,p}(\Omega)\big)'$, the set of bounded linear functionals on $W_0^{1,p}(\Omega)$, see Example 3.2 below for more details. In the present paper, we take for the intrinsic distance of u_h to u the *dual norm* of the residual $\mathcal{R}(u_h)$

$$\|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)} := \sup_{v \in W_0^{1,p}(\Omega);\, \|\nabla v\|_p = 1} \langle \mathcal{R}(u_h), v\rangle_{W^{-1,q}(\Omega), W_0^{1,p}(\Omega)}; \tag{1.6}$$

of course $\|\mathcal{R}(u_h)\|_{W^{-1,2}(\Omega)} = \|\nabla(u - u_h)\|$ when $p = 2$ and $\boldsymbol{\sigma}(\boldsymbol{g}) = \boldsymbol{g}$. Note, however, that other distances might be called intrinsic. Considering for simplicity $u^{\mathrm{D}} = 0$ and defining the energy by

$$\mathcal{E}(v) := \frac{1}{p}\|\nabla v\|_p^p - (f, v) = \int_\Omega \left(\frac{1}{p}|\nabla v|^p - fv\right) \mathrm{d}\boldsymbol{x} \qquad v \in W_0^{1,p}(\Omega), \tag{1.7}$$

the energy difference $\mathcal{E}(u_h) - \mathcal{E}(u)$ is often considered as the intrinsic distance, see, e.g., Repin [46, Section 8.4.1], and is actually proportional to the squared quasi-norm error (that again can be used as an intrinsic distance) introduced by Barrett and Liu in [6, 7], see Diening and Kreuzer [29, Lemma 16] or Belenki *et al.* [10, Lemma 3.2], cf. also Remark 3.5 below.

Sticking to (1.6), the analog of (1.3) can then be obtained: there are a posteriori estimators $\eta_K(u_h)$, fully and locally computable from u_h, such that

$$\|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)} \le \left\{ \sum_{K \in \mathcal{T}_h} \eta_K(u_h)^q \right\}^{\frac{1}{q}}, \tag{1.8}$$

see, e.g., Verfürth [52, 53], Veeser and Verfürth [50], El Alaoui *et al.* [31], Ern and Vohralík [33], or Kreuzer and Süli [40]. This can typically be proved under the *orthogonality condition*

$$\langle \mathcal{R}(u_h), \psi_{\boldsymbol{a}}\rangle_{W^{-1,q}(\Omega), W_0^{1,p}(\Omega)} = 0 \qquad \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{1.9}$$

where $\mathcal{V}_h^{\mathrm{int}}$ stands for interior vertices of the mesh $\mathcal{T}_h$ and $\psi_{\boldsymbol{a}}$ are test functions forming a partition of unity over all vertices $\boldsymbol{a} \in \mathcal{V}_h$. However, the analog of the local efficiency (1.4) for $p \neq 2$ does not seem to be obvious. The foremost reason is that the intrinsic dual error measure (1.6) does not seem to be localizable in the sense that

$$\|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)} \neq \left\{ \sum_{K \in \mathcal{T}_h} \|\mathcal{R}(u_h)\|_{W^{-1,q}(K)}^q \right\}^{\frac{1}{q}},$$

in contrast to (1.2).

For certain estimators from (1.8) with piecewise polynomial u_h, global efficiency in the form

$$\left\{ \sum_{K \in \mathcal{T}_h} \eta_K(u_h)^q \right\}^{\frac{1}{q}} \le C \|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)}$$

has been shown previously, cf. [53, 31, 33, 40] and the references therein. Carrying on the results and the proofs in [52, 53], it is, in fact, possible to show that

$$\eta_K(u_h) \leq C \left\{ \sum_{\boldsymbol{a} \in \mathcal{V}_K} \|\mathcal{R}(u_h)\|^q_{W^{-1,q}(\omega_{\boldsymbol{a}})} \right\}^{\frac{1}{q}}, \tag{1.10}$$

where $\mathcal{V}_K$ stands for the vertices of the element $K \in \mathcal{T}_h$ and $\omega_{\boldsymbol{a}}$ is a patch of mesh elements around the vertex $\boldsymbol{a}$, see for example [33, Theorem 5.3], [31, proof of Lemma 4.3], [40, proof of Theorem 21] for general $p \in (1, \infty)$, or [28, equation (3.10)] for the Hilbertian setting $p = 2$. Note, however, that all these results are connected with a certain class of PDE problems considered in these studies as well as with a certain appropriately constructed error estimator. To conclude, we observe the following points:

1. Inequality (1.8) together with (1.10) imply

$$\|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)} \leq C_1 \left\{ \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}(u_h)\|^q_{W^{-1,q}(\omega_{\boldsymbol{a}})} \right\}^{\frac{1}{q}}. \tag{1.11a}$$

For $p = q = 2$, this has probably been first shown in Babuška and Miller [5, Theorem 2.1.1].

2. It can also be shown that

$$\left\{ \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}(u_h)\|^q_{W^{-1,q}(\omega_{\boldsymbol{a}})} \right\}^{\frac{1}{q}} \leq C_2 \|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)}. \tag{1.11b}$$

See in particular [5, Theorem 2.1.1], Cohen *et al.* [28, equation (3.23)], Ciarlet and Vohralík [27, Theorem 3.3], and the revised version of Ern and Guermond [32] for $p = q = 2$.

3. Thus, for the error measure $\|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)}$, the a posteriori estimators $\eta_K(u_h)$ lead to an a posteriori analysis framework where one has localization of the error measure (1.11), global reliability (1.8), and local efficiency (1.10). This is thus a fully consistent and analogous situation to (1.2), (1.3), and (1.4) of the $W^{1,2}_0(\Omega)$ setting.

The main purpose of the present paper is to give a minimalist and direct proof of the two inequalities (1.11) for general exponent p, including also the limiting cases $p = 1$ and $p = \infty$, and without considering any particular partial differential equation or a posteriori error estimators. In particular, Theorem 3.7 shows that, under the orthogonality condition (1.9), *dual norms of all functionals in* $W^{-1,q}(\Omega)$ are *localizable* in the sense that (1.11) holds, with C_1 and C_2 *only depending* on the *regularity* of the partition $\mathcal{T}_h$; in particular the constants are robust with respect to exponent $p \in [1, \infty]$. The orthogonality condition (1.9) is only necessary for robustness of C_1 with respect to the mesh size h; the constant C_2 depends merely on maximal overlap of the partition $\cup_{\boldsymbol{a} \in \mathcal{V}_h} \omega_{\boldsymbol{a}}$. The result of Theorem 3.7 applies to, but is not limited to, dual norms of residuals of (nonlinear) partial differential equations of the form (1.5). Besides implying local a posteriori error efficiency, the localization of a seemingly only global distance of the form (1.11) may have important consequences in the adaptive approximation theory or for equivalence of local-best and global-best approximations as in [49]. We discuss localization of the $W^{1,p}_0(\Omega)$-norm error in Remark 3.4 and the localization of the global lifting of $\mathcal{R}(u_h)$ into $W^{1,p}_0(\Omega)$ in Remark 3.6. Remark 3.10 further shows that (1.11b) can be strengthened to hold patch by patch $\omega_{\boldsymbol{a}}$, with a global lifting of $\mathcal{R}(u_h)$ on the right-hand side. All these results are presented in Section 3, after we set up the notation and gather the preliminaries in Section 2.

Localization concepts that take form similar to (1.11) also appear in the theory of function spaces, cf. Triebel [48], where they are of independent interest. Consider the Whitney covering of the domain Ω, which we here denote as $\{\omega_{\boldsymbol{a}}\}_{\boldsymbol{a} \in \mathbb{N}}$, and a subordinate partition of unity $\sum_{\boldsymbol{a} \in \mathbb{N}} \psi_{\boldsymbol{a}} = 1$ with $0 \leq \psi_{\boldsymbol{a}} \leq 1$, $\psi_{\boldsymbol{a}}$ smooth, compactly supported in $\omega_{\boldsymbol{a}}$, and all derivatives controlled by distance to boundary: $\|\nabla^M \psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}} \le C_M \operatorname{dist}(\omega_{\boldsymbol{a}}, \partial\Omega)^{-M}$ for all $M \in \mathbb{N}$. For E-thick domains Ω (which includes bounded Lipschitz domains), there holds the so-called *refined localization property*

$$\|v\|_{W^{-1,q}(\Omega)} \approx \left\{ \sum_{\boldsymbol{a}\in\mathbb{N}} \|\psi_{\boldsymbol{a}} v\|^q_{W^{-1,q}(\omega_{\boldsymbol{a}})} \right\}^{\frac{1}{q}} \qquad \forall v \in W^{-1,q}(\Omega), \tag{1.12}$$

i.e., the term on the right-hand side is an equivalent quasi-norm. For precise definitions and statements, see [48, Theorem 3.28]. This result holds for spaces of F-scale comprising Lizorkin–Triebel and classical Sobolev spaces, including negative differentiability and specially the case $W^{-1,q}(\Omega)$, which is incidentally of interest here and which we only indicate in (1.12). Note that there is no sequence of partitions here (the partition $\{\omega_{\boldsymbol{a}}\}_{\boldsymbol{a}\in\mathbb{N}}$ is fixed, arbitrarily fine close to the boundary $\partial\Omega$). In contrast, the aim of this study is robustness of the constants C_1 and C_2 in (1.11) with respect to all possible partitions $\mathcal{T}_h$ (subject only to regularity), including arbitrary refinement in the interior of the domain Ω.

Finally, we are also interested in the situations where the orthogonality condition (1.9) is not satisfied. In practical applications, this is typically connected with inexact algebraic/nonlinear solvers. Our Theorems 4.1 and 4.3 give two-sided bounds on $\|\mathcal{R}(u_h)\|_{W^{-1,q}(\Omega)}$ in this setting and Corollary 4.8 proves therefrom that the h- and *p-robust* localization result of Theorem 3.7 can be recovered provided that the loss of orthogonality is small with respect to the leading term. In Remark 4.2, we comment that (1.11) holds even without orthogonality condition (1.9), but with C_1 deteriorating with mesh refinement (for decreasing h). This is intuitively consistent with the result (1.12), where the fixed partition is coarse in the interior of Ω and arbitrarily fine only close to the boundary $\partial\Omega$. We collect these results in Section 4, including Theorem 4.10 that presents an extension to vectorial setting. Its typical practical applications stem from fluid mechanics or elasticity; the results established here indeed represent one of the key tools used in [14] for deriving a complete theory of a posteriori error estimation for implicit constitutive relations in the generalized Stokes setting, capturing the most common nonlinear fluid models in a unified way. To conclude, Section 5 illustrates our theoretical findings on several numerical experiments.

2 Setting

We describe the setting and notation in this section, detailing the partition of unity that will be central in our developments. We then state cut-off estimates based on Poincaré–Friedrichs inequalities necessary later.

2.1 Notation, assumptions, and a partition of unity

We suppose that $\Omega \subset \mathbb{R}^d$, $d \ge 1$, is a domain (open, bounded, and connected set) with a Lipschitz-continuous boundary and diameter h_Ω. Let $1 \le p \le \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. We will work with standard Sobolev spaces $W^{1,p}(\Omega)$ of functions with $L^p(\Omega)$-integrable weak derivatives, see, e.g., Evans [35], Brenner and Scott [16], and the references therein. The space $W^{1,p}_0(\Omega)$ then stands for functions that are zero in the sense of traces on $\partial\Omega$. Similar notation is used on subdomains of Ω.

For measurable subset $\omega \subset \Omega$ and functions $u \in L^q(\omega)$, $v \in L^p(\omega)$, $(u, v)_\omega$ stands for $\int_\omega u(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and similarly $(\boldsymbol{u}, \boldsymbol{v})_\omega := \int_\omega \boldsymbol{u}(\boldsymbol{x})\cdot\boldsymbol{v}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for $\boldsymbol{u} \in [L^q(\omega)]^d$ and $\boldsymbol{v} \in [L^p(\omega)]^d$; we simply write (u, v) instead of $(u, v)_\Omega$ when $\omega = \Omega$ and similarly in the vectorial case. We follow the convention $\|v\|_{p,\omega} := \left(\int_\omega |v(\boldsymbol{x})|^p\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}}$ for $1 \le p < \infty$, $\|v\|_{\infty,\omega} := \operatorname{ess\,sup}_{\boldsymbol{x}\in\omega} |v(\boldsymbol{x})|$, $\|\boldsymbol{v}\|_{p,\omega} := \left(\int_\omega |\boldsymbol{v}(\boldsymbol{x})|^p\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}}$ for $1 \le p < \infty$, and $\|\boldsymbol{v}\|_{\infty,\omega} := \operatorname{ess\,sup}_{\boldsymbol{x}\in\omega} |\boldsymbol{v}(\boldsymbol{x})|$, where $|\boldsymbol{v}| = \left(\sum_{i=1}^d |\boldsymbol{v}_i|^2\right)^{\frac{1}{2}}$ is the Euclidean norm in $\mathbb{R}^d$. Note that, when $p \ne 2$, $\|\nabla v\|_{p,\omega}$ is different from (but equivalent to) $|v|_{1,p,\omega} = \left(\sum_{i=1}^d \|\partial_{\boldsymbol{x}_i} v\|^p_{p,\omega}\right)^{\frac{1}{p}}$ if $1 \le p < \infty$, $|v|_{1,\infty,\omega} = \max_{i=1,\dots,d} \|\partial_{\boldsymbol{x}_i} v\|_{\infty,\omega}$ for $v \in W^{1,p}(\omega)$; we will often use below the equivalence of $l^p(\mathbb{R}^m)$ norms $|\mathrm{v}|_p := \left(\sum_{i=1}^m |\mathrm{v}_i|^p\right)^{\frac{1}{p}}$

if $1 \le p < \infty$, $|\mathrm{v}|_\infty := \max_{i=1,\ldots,m} |\mathrm{v}_i|$

$$|\mathrm{v}|_p \le |\mathrm{v}|_q \le m^{\frac{1}{q}-\frac{1}{p}} |\mathrm{v}|_p \qquad \forall \mathrm{v} \in \mathbb{R}^m,\ 1 \le q \le p \le \infty. \tag{2.1}$$

We also denote by $|\cdot|_2$ the spectral matrix norm, given by $|\mathbb{A}|_2 := \max_{\mathrm{v}\in\mathbb{R}^m;\ |\mathrm{v}|_2=1} |\mathbb{A}\mathrm{v}|_2$ for a matrix $\mathbb{A} \in \mathbb{R}^{m\times m}$.

We suppose that there exists a partition of unity

$$\sum_{\boldsymbol{a}\in\mathcal{V}_h} \psi_{\boldsymbol{a}} = 1 \quad \text{ a.e. in } \Omega \tag{2.2}$$

by functions $\psi_{\boldsymbol{a}} \in W^{1,\infty}(\Omega)$ with a local support denoted by $\overline{\omega_{\boldsymbol{a}}}$. More precisely, $\omega_{\boldsymbol{a}}$ are open subdomains of the domain Ω of nonzero d-dimensional measure, diameter $h_{\omega_{\boldsymbol{a}}}$, with a Lipschitz-continuous boundary, and satisfying $\cup_{\boldsymbol{a}\in\mathcal{V}_h}\omega_{\boldsymbol{a}} = \Omega$; $\omega_{\boldsymbol{a}}$ is called a *patch*. The index $\boldsymbol{a}$ denotes a point in $\overline{\omega_{\boldsymbol{a}}}$ called a *vertex*, termed interior if $\boldsymbol{a} \in \Omega$ and termed boundary if $\boldsymbol{a} \in \partial\Omega$; the corresponding index sets are $\mathcal{V}_h = \mathcal{V}_h^{\mathrm{int}} \cup \mathcal{V}_h^{\mathrm{ext}}$, $\mathcal{V}_h^{\mathrm{int}} \cap \mathcal{V}_h^{\mathrm{ext}} = \emptyset$. For $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}$, $\partial\omega_{\boldsymbol{a}} \cap \partial\Omega$ is supposed to have a nonzero $(d-1)$-dimensional measure. We identify $\psi_{\boldsymbol{a}}$ with $\psi_{\boldsymbol{a}}|_{\omega_{\boldsymbol{a}}}$ and suppose that $\psi_{\boldsymbol{a}}$ takes values between 0 and 1 on $\omega_{\boldsymbol{a}}$, $\|\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}} \le 1$; $\psi_{\boldsymbol{a}}$ is zero in the sense of traces on the whole boundary $\partial\omega_{\boldsymbol{a}}$ for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$ and on $\partial\omega_{\boldsymbol{a}} \setminus \partial\Omega$ for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}$.

The partition of the domain Ω by the patches $\omega_{\boldsymbol{a}}$ needs to be overlapping, i.e., the intersection of several different patches has a nonzero d-dimensional measure. We collect the closures of the minimal intersections into a nonoverlapping partition $\mathcal{T}_h$ of Ω with closed *elements* denoted by K, with diameter h_K. We suppose that each point in Ω lies in at most N_{ov} patches. Equivalently, each $K \in \mathcal{T}_h$ corresponds to the closure of intersection of at most N_{ov} patches, and we collect their vertices $\boldsymbol{a}$ in the set $\mathcal{V}_K$. Vice-versa, the elements $K \in \mathcal{T}_{\boldsymbol{a}}$ cover $\overline{\omega_{\boldsymbol{a}}}$. There in particular holds

$$\left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|v\|_{p,\omega_{\boldsymbol{a}}}^p \right\}^{\frac{1}{p}} \le \|v\|_p \qquad \forall v \in L^p(\Omega),\ 1 \le p < \infty, \tag{2.3a}$$

$$\max_{\boldsymbol{a}\in\mathcal{V}_h} \|v\|_{\infty,\omega_{\boldsymbol{a}}} = \|v\|_\infty \qquad \forall v \in L^\infty(\Omega). \tag{2.3b}$$

We shall frequently use the patchwise Sobolev spaces given by

$$W_*^{1,p}(\omega_{\boldsymbol{a}}) := \begin{cases} \{v \in W^{1,p}(\omega_{\boldsymbol{a}});\ (v,1)_{\omega_{\boldsymbol{a}}} = 0\} & \text{if } \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}, \\ \{v \in W^{1,p}(\omega_{\boldsymbol{a}});\ v = 0 \text{ on } \partial\omega_{\boldsymbol{a}} \cap \partial\Omega\} & \text{if } \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}, \end{cases} \tag{2.4}$$

having zero mean value over $\omega_{\boldsymbol{a}}$ in the first case and vanishing trace on the boundary of Ω in the second case. The Poincaré–Friedrichs inequality then states that

$$\|v\|_{p,\omega_{\boldsymbol{a}}} \le C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} h_{\omega_{\boldsymbol{a}}} \|\nabla v\|_{p,\omega_{\boldsymbol{a}}} \qquad \forall v \in W_*^{1,p}(\omega_{\boldsymbol{a}}), \tag{2.5}$$

where, recall, $h_{\omega_{\boldsymbol{a}}}$ stands for the diameter of the patch $\omega_{\boldsymbol{a}}$. In particular, for $1 < p < \infty$, $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$, and $\omega_{\boldsymbol{a}}$ convex, $C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} = 2\left(\frac{p}{2}\right)^{\frac{1}{p}}$, see Chua and Wheeden [26]; for $p = 1$, $C_{\mathrm{PF},1,\omega_{\boldsymbol{a}}} = \frac{1}{2}$ in this setting, see Acosta and Durán [1] or [26], and for $p = \infty$, $C_{\mathrm{PF},\infty,\omega_{\boldsymbol{a}}} = 1$ is straightforward. This implies that $\frac{1}{2} \le C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} \le C_{\mathrm{PF},2\mathrm{e},\omega_{\boldsymbol{a}}} = 2\,\mathrm{e}^{\frac{1}{2\mathrm{e}}} \approx 2.404$ for all $1 \le p \le \infty$ and a convex interior patch. The values for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$ and nonconvex patches $\omega_{\boldsymbol{a}}$ are identified in, e.g., Veeser and Verfürth [51, Theorems 3.1 and 3.2] for $1 \le p < \infty$; whenever $\omega_{\boldsymbol{a}}$ is star-shaped, $C_{\mathrm{PF},\infty,\omega_{\boldsymbol{a}}} = 2$. Finally, $C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} = 1$ for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}$ when $\partial\omega_{\boldsymbol{a}} \cap \partial\Omega$ can be reached in a constant direction from any point inside $\omega_{\boldsymbol{a}}$; bounds in the general case can be obtained for instance as in [50, Lemma 5.1]. We describe the *regularity* of the partition by the number

$$C_{\mathrm{cont,PF}} := \max_{\boldsymbol{a}\in\mathcal{V}_h} \{1 + C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} h_{\omega_{\boldsymbol{a}}} \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}\}, \tag{2.6}$$

which we suppose to be uniformly bounded on families of the considered partitions indexed by the parameter $h := \max_{\boldsymbol{a}\in\mathcal{V}_h} h_{\omega_{\boldsymbol{a}}}$.

2.2 Examples of partitions of unity

We now give three examples of possible partitions of unity $\psi_{\boldsymbol{a}}$ and subdomains $\omega_{\boldsymbol{a}}$.

Example 2.1 (Simplicial or parallelepipedal meshes from the finite element context)**.** *A prototypal example we have in mind is the case where* Ω *is a polytope,* $\cup_{K\in\mathcal{T}_h} K = \overline{\Omega}$*, each element* K *is a closed* d*-dimensional simplex (triangle in* $d = 2$*, tetrahedron in* $d = 3$*) or a* d*-dimensional parallelepiped (quadrilateral in* $d = 2$*, hexahedron in* $d = 3$*), and the intersection of two different elements* K *is either empty or their* d'*-dimensional common face,* $0 \leq d' \leq d-1$*. Then* $N_{\mathrm{ov}} = d+1$ *for simplices and* $N_{\mathrm{ov}} = 2^d$ *for parallelepipeds,* $\omega_{\boldsymbol{a}}$ *is the patch of all elements sharing the given vertex* $\boldsymbol{a} \in \mathcal{V}_h$*, and* (2.3a) *takes form of equality. In particular, for the seminorm on* $W^{1,p}(\Omega)$*,*

$$\frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla v\|_{p,\omega_{\boldsymbol{a}}}^p = \|\nabla v\|_p^p \qquad \forall v \in W^{1,p}(\Omega),\ 1 \leq p < \infty, \tag{2.7a}$$

$$\max_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla v\|_{\infty,\omega_{\boldsymbol{a}}} = \|\nabla v\|_\infty \qquad \forall v \in W^{1,\infty}(\Omega). \tag{2.7b}$$

We then take $\psi_{\boldsymbol{a}}$ *as the continuous, piecewise (d-)affine "hat" function of finite element analysis, taking value* 1 *at the vertex* $\boldsymbol{a} \in \mathcal{V}_h$ *and* 0 *in all other vertices from* $\mathcal{V}_h$*. Denoting by* $\kappa_{\mathcal{T}_h}$ *the mesh shape-regularity parameter given by the maximal ratio of the diameter of* K *to the diameter of the largest ball inscribed into* K *over all* $K \in \mathcal{T}_h$*, it follows from Veeser and Verfürth [51, Theorems 3.1 and 3.2], Carstensen and Funken [23], or Braess* et al. *[15] that both* $C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}}$ *of* (2.5) *and* $C_{\mathrm{cont,PF}}$ *of* (2.6) *only depend on* $\kappa_{\mathcal{T}_h}$*. Note further that in the context of approximation of the solution of a partial differential equation by the finite element method, with the residual* $\mathcal{R}$ *described in Remark 3.2 below, the crucial orthogonality condition* (3.20) *amounts to requesting the presence of the hat functions* $\psi_{\boldsymbol{a}}$*,* $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$*, in the finite element basis.*

Example 2.2 (B-splines supports from the isogeometric analysis context)**.** *Let the space dimension* $d = 1$*, let* Ω *be an interval, and let* $\mathcal{T}_h$ *be a mesh of* Ω *consisting of intervals* K *of size* h_K*,* $\cup_{K\in\mathcal{T}_h} K = \overline{\Omega}$*. B-splines are non-negative piecewise (with respect to* $\mathcal{T}_h$*) polynomials of degree* k *and class* C^l*,* $k \geq 1$*,* $0 \leq l \leq k-1$*, with smallest possible support and given scaling; typically* $l = k-1$*, i.e., one requests continuity of the derivatives up to order* $k-1$*. Denoting them* $\psi_{\boldsymbol{a}}$*, the subdomains* $\omega_{\boldsymbol{a}}$ *can simply be taken as the supports of the B-splines* $\psi_{\boldsymbol{a}}$*. Then the vertices* $\boldsymbol{a}$ *that form the set* $\mathcal{V}_h$ *lie inside* $\omega_{\boldsymbol{a}}$ *if the value of* $\psi_{\boldsymbol{a}}$ *on the boundary of the domain* Ω *is zero, and are the corresponding endpoints of* Ω *otherwise. Crucially, the partition of unity* (2.2) *holds for B-splines. For* $k = 1$ *and* $l = 0$ *(piecewise affine functions with* C^0 *continuity), this setting coincides with the finite element context of Remark 2.1. In general, however, the subdomains* $\omega_{\boldsymbol{a}}$ *are larger here, leading to increased overlap between* $\omega_{\boldsymbol{a}}$ *and higher value of the overlap parameter* N_{ov}*, in dependence on the continuity parameter* l*. In the context of the partial differential equation residual* $\mathcal{R}$ *of Remark 3.2 below, the orthogonality condition* (3.20) *amounts to the use of the B-splines/isogeometric analysis approximation, see Bazilevs* et al. *[8] or Buffa and Giannelli [18] and the references therein. Extension to higher space dimensions* $d > 1$ *is straightforward by tensor products for* Ω *being a rectangular parallelepiped; general domains can be treated via non-uniform rational basis splines (NURBS) and transformation from the parametric space into the physical space.*

Example 2.3 (Meshfree methods)**.** *In general, the approach developed here can be applied to any setting that is based on the idea of basis functions having local (small, compact) support and forming the partition of unity* (2.2)*. The partition of unity method, see Babuška and Melenk [43, 4], and in general meshfree methods, see [37] and the references therein, can serve as examples.*

2.3 Poincaré–Friedrichs cut-off estimates

The forthcoming result, following the lines of Carstensen and Funken [23, Theorem 3.1] or Braess *et al.* [15, Section 3], with $W^{1,p}(\omega_{\boldsymbol{a}})$-Poincaré–Friedrichs inequalities of Chua and Wheeden [26] and Veeser and Verfürth [51], will form the basic building block for our considerations:

Lemma 2.4 (Cut-off estimate)**.** *For the constant* $C_{\mathrm{cont,PF}}$ *from* (2.6)*, there holds, for all* $\boldsymbol{a} \in \mathcal{V}_h$*,*

$$\|\nabla(\psi_{\boldsymbol{a}} v)\|_{p,\omega_{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF}} \|\nabla v\|_{p,\omega_{\boldsymbol{a}}} \qquad \forall v \in W_*^{1,p}(\omega_{\boldsymbol{a}}),\ 1 \leq p \leq \infty.$$

Proof. Let $\boldsymbol{a} \in \mathcal{V}_h$. We have, employing the triangle inequality, $\|\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}} = 1$, and (2.5),

$$\begin{aligned}\|\nabla(\psi_{\boldsymbol{a}} v)\|_{p,\omega_{\boldsymbol{a}}} &= \|\nabla\psi_{\boldsymbol{a}} v + \psi_{\boldsymbol{a}}\nabla v\|_{p,\omega_{\boldsymbol{a}}}\\ &\le \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}\|v\|_{p,\omega_{\boldsymbol{a}}} + \|\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}\|\nabla v\|_{p,\omega_{\boldsymbol{a}}}\\ &\le (1 + C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} h_{\omega_{\boldsymbol{a}}}\|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}})\|\nabla v\|_{p,\omega_{\boldsymbol{a}}},\end{aligned}$$

and the assertion follows from the definition (2.6). □

2.4 An overlapping-patches estimate

We finally provide an auxiliary coloring-type estimate for a sum of functions from $W_0^{1,p}(\omega_{\boldsymbol{a}})$ that will be used later.

Lemma 2.5 (An overlapping-patches estimate)**.** *Let $1 \le p \le \infty$. Assume there is a collection of functions $\{v^{\boldsymbol{a}}\}_{\boldsymbol{a}\in\mathcal{V}_h}$ with $v^{\boldsymbol{a}} \in W_0^{1,p}(\omega_{\boldsymbol{a}})$, extended by zero to $W_0^{1,p}(\Omega)$. Then $\sum_{\boldsymbol{a}\in\mathcal{V}_h} v^{\boldsymbol{a}} \in W_0^{1,p}(\Omega)$ and*

$$\left\|\nabla\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h} v^{\boldsymbol{a}}\right\|_p \le \left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\nabla v^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}^p\right\}^{\frac{1}{p}} \qquad \textit{if } 1 \le p < \infty, \tag{2.8a}$$

$$\left\|\nabla\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h} v^{\boldsymbol{a}}\right\|_\infty \le \max_{\boldsymbol{a}\in\mathcal{V}_h}\|\nabla v^{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}} \qquad \textit{if } p = \infty. \tag{2.8b}$$

Proof. Let $1 \le p < \infty$. Fix $K \in \mathcal{T}_h$ and denote the number of vertices in K by $|\mathcal{V}_K|$. The triangle and Hölder inequalities give, a.e. in K,

$$\left|\sum_{\boldsymbol{a}\in\mathcal{V}_K}\nabla v^{\boldsymbol{a}}\right| \le \sum_{\boldsymbol{a}\in\mathcal{V}_K}|\nabla v^{\boldsymbol{a}}| \le \left\{\sum_{\boldsymbol{a}\in\mathcal{V}_K}|\nabla v^{\boldsymbol{a}}|^p\right\}^{\frac{1}{p}}|\mathcal{V}_K|^{\frac{1}{q}}.$$

By assumption it holds $|\mathcal{V}_K| \le N_{\mathrm{ov}}$ for every $K \in \mathcal{T}_h$, so that, since $\frac{p}{q} = p - 1$,

$$\left|\nabla\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_K} v^{\boldsymbol{a}}\right|^p \le \frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_K}|\nabla v^{\boldsymbol{a}}|^p. \tag{2.9}$$

Integrating both sides of (2.9) over K, summing over all $K \in \mathcal{T}_h$, and taking $\frac{1}{p}$-th power of the result gives (2.8a). Estimate (2.8b) is trivial. □

3 Localization of dual functional norms

In this section we state and prove our main localization result; we also present some of its consequences. We first fix some notations and introduce the overall context in more detail.

3.1 Context

For given $p \in [1,\infty]$, denote

$$V := W_0^{1,p}(\Omega) \tag{3.1}$$

and consider a bounded linear functional $\mathcal{R} \in V'$. We denote the action of $\mathcal{R}$ on $v \in V$ by $\langle\mathcal{R}, v\rangle_{V',V}$ and define

$$\|\mathcal{R}\|_{V'} := \sup_{v\in V;\,\|\nabla v\|_p=1}\langle\mathcal{R}, v\rangle_{V',V}. \tag{3.2}$$

Similarly, for vertex $\boldsymbol{a} \in \mathcal{V}_h$ and the corresponding patch subdomain $\omega_{\boldsymbol{a}}$, set

$$V^{\boldsymbol{a}} := W_0^{1,p}(\omega_{\boldsymbol{a}})$$

and define the restriction of the functional $\mathcal{R}$ to $V^{\boldsymbol{a}}$, still denoted by $\mathcal{R}$, via

$$\langle\mathcal{R}, v\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}} := \langle\mathcal{R}, v\rangle_{V',V} \qquad v \in V^{\boldsymbol{a}}, \tag{3.3}$$

where $v \in V^{\boldsymbol{a}}$ is extended by zero outside of the patch $\omega_{\boldsymbol{a}}$ to $v \in V$. Let

$$\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} := \sup_{v\in V^{\boldsymbol{a}};\,\|\nabla v\|_{p,\omega_{\boldsymbol{a}}}=1}\langle\mathcal{R}, v\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}}. \tag{3.4}$$

3.2 Examples of functionals $\mathcal{R}$

To fix ideas, we give two examples fitting in the context of Section 3.1.

Example 3.1 ($\mathcal{R}$ being divergence of an integrable function)**.** *Let $\boldsymbol{\xi} \in [L^q(\Omega)]^d$. A simple example of $\mathcal{R} \in V'$ is*

$$\langle \mathcal{R}, v\rangle_{V',V} := (\boldsymbol{\xi}, \nabla v) \qquad v \in V. \tag{3.5}$$

In this case, immediately, for any $\boldsymbol{a} \in \mathcal{V}_h$,

$$\langle \mathcal{R}, v\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}} = (\boldsymbol{\xi}, \nabla v)_{\omega_{\boldsymbol{a}}} \qquad v \in V^{\boldsymbol{a}}.$$

Moreover, using definitions (3.2) *and* (3.4)*, we easily obtain via the Hölder inequality the bounds*

$$\|\mathcal{R}\|_{V'} \le \|\boldsymbol{\xi}\|_q, \tag{3.6a}$$

$$\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \le \|\boldsymbol{\xi}\|_{q,\omega_{\boldsymbol{a}}} \qquad \forall \boldsymbol{a} \in \mathcal{V}_h. \tag{3.6b}$$

Example 3.2 ($\mathcal{R}$ given by a residual of a partial differential equation)**.** *Let $1 \le p \le \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, $u^{\mathrm{D}} \in W^{1,p}(\Omega)$, $f \in L^q(\Omega)$, and let $(u, \boldsymbol{\sigma})$ be a weak solution*[2] *to the problem*

$$-\operatorname{div} \boldsymbol{\sigma} = f \qquad \text{in } \Omega, \tag{3.9a}$$

$$u = u^{\mathrm{D}} \qquad \text{on } \partial\Omega, \tag{3.9b}$$

$$\boldsymbol{h}(\boldsymbol{\sigma}, \nabla u) = \boldsymbol{0} \qquad \text{in } \Omega. \tag{3.9c}$$

Here $\boldsymbol{\sigma} \in [L^q(\Omega)]^d$, $u \in W^{1,p}(\Omega)$ such that $u - u^{\mathrm{D}} \in W_0^{1,p}(\Omega)$, and a nonlinear function $\boldsymbol{h} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is such that it holds, with some $\alpha, \beta > 0$,

$$\boldsymbol{s}\cdot\boldsymbol{d} \ge \alpha\big(|\boldsymbol{s}|^q + |\boldsymbol{d}|^p\big) - \beta \qquad \text{whenever } \boldsymbol{s}, \boldsymbol{d} \in \mathbb{R}^d \text{ and } \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{0}. \tag{3.10}$$

[2] Assuming $1 < p < \infty$, weak solution to problem (3.9) can be defined as: to find $u : \Omega \to \mathbb{R}$ and $\boldsymbol{\sigma} : \Omega \to \mathbb{R}^d$ such that

$$u - u^{\mathrm{D}} \in V, \tag{3.7a}$$

$$\boldsymbol{\sigma} \in [L^q(\Omega)]^d, \tag{3.7b}$$

$$(\boldsymbol{\sigma}, \nabla v) = (f, v) \qquad \forall v \in V, \tag{3.7c}$$

$$\boldsymbol{h}(\boldsymbol{\sigma}, \nabla u) = \boldsymbol{0} \qquad \text{almost everywhere in } \Omega. \tag{3.7d}$$

This problem has at least one solution if, for example, the function $\boldsymbol{h}$ fulfills, with some $\alpha, \beta > 0$, the following conditions:

1. $\boldsymbol{h}(\boldsymbol{0}, \boldsymbol{0}) = \boldsymbol{0}$;
2. if $\boldsymbol{s}^1, \boldsymbol{s}^2, \boldsymbol{d}^1, \boldsymbol{d}^2 \in \mathbb{R}^d$ and $\boldsymbol{h}(\boldsymbol{s}^1, \boldsymbol{d}^1) = \boldsymbol{h}(\boldsymbol{s}^2, \boldsymbol{d}^2) = \boldsymbol{0}$ then
$$\big(\boldsymbol{s}^1 - \boldsymbol{s}^2\big)\cdot\big(\boldsymbol{d}^1 - \boldsymbol{d}^2\big) \ge 0; \tag{3.8}$$
3. if the couple $(\boldsymbol{s}, \boldsymbol{d}) \in \mathbb{R}^d \times \mathbb{R}^d$ fulfills
$$\big(\boldsymbol{s} - \tilde{\boldsymbol{s}}\big)\cdot\big(\boldsymbol{d} - \tilde{\boldsymbol{d}}\big) \ge 0 \qquad \text{for all } \tilde{\boldsymbol{s}}, \tilde{\boldsymbol{d}} \in \mathbb{R}^d \text{ with } \boldsymbol{h}(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{d}}) = \boldsymbol{0},$$
then $(\boldsymbol{s}, \boldsymbol{d})$ also fulfills $\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{0}$;
4. if $\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{0}$ then (3.10) holds;

see [36] and also [19, 21, 20] for fluid mechanics context. If, in addition, inequality (3.8) is strict whenever $\boldsymbol{s}^1 \ne \boldsymbol{s}^2$ and $\boldsymbol{d}^1 \ne \boldsymbol{d}^2$, then such a solution is unique.

For a novel theory of weak solutions in the non-reflexive case $p = \infty$ and within the context of solid mechanics, we refer the interested reader to [9].

Typical examples of function $\boldsymbol{h}$ are

$$\left.\begin{aligned}\boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{s} - \left(1+|\boldsymbol{d}|^2\right)^{\frac{p-2}{2}}\boldsymbol{d},\\ \boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{s} - \left(1+|\boldsymbol{d}|\right)^{p-2}\boldsymbol{d},\end{aligned}\right\}\ \textit{regularized p-Laplace}$$

$$\left.\begin{aligned}\boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{d} - \left(1+|\boldsymbol{s}|^2\right)^{\frac{q-2}{2}}\boldsymbol{s},\\ \boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{d} - \left(1+|\boldsymbol{s}|\right)^{q-2}\boldsymbol{s},\end{aligned}\right\}\ \textit{generalized p-Laplace}$$

$$\left.\begin{aligned}\boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{d} - \frac{(|\boldsymbol{s}|-\sigma_*)^+}{|\boldsymbol{s}|}\left(1+|\boldsymbol{s}|^2\right)^{\frac{q-2}{2}}\boldsymbol{s},\\ \boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{d} - \frac{(|\boldsymbol{s}|-\sigma_*)^+}{|\boldsymbol{s}|}\left(1+|\boldsymbol{d}|^2\right)^{-\frac{p-2}{2}}\boldsymbol{s},\\ \boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{s} - \frac{(|\boldsymbol{d}|-\delta_*)^+}{|\boldsymbol{d}|}\left(1+|\boldsymbol{d}|^2\right)^{\frac{p-2}{2}}\boldsymbol{d},\end{aligned}\right\}\ \textit{activated p-Laplace}$$

$$\left.\begin{aligned}\boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{s} - |\boldsymbol{d}|^{p-2}\boldsymbol{d},\\ \boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{d} - |\boldsymbol{s}|^{q-2}\boldsymbol{s},\end{aligned}\right\}\ \textit{classical p-Laplace}$$

where $(t)^+ = \max(t,0)$ and $\sigma_*, \delta_* \geq 0$ are given real parameters. Note that the last two examples give identical response since

$$\boldsymbol{s} = |\boldsymbol{d}|^{p-2}\boldsymbol{d} \qquad \Longleftrightarrow \qquad \boldsymbol{d} = |\boldsymbol{s}|^{q-2}\boldsymbol{s}.$$

Consequently

$$\boldsymbol{s}\cdot\boldsymbol{d} = \left(\frac{1}{p}+\frac{1}{q}\right)\boldsymbol{s}\cdot\boldsymbol{d} = \frac{|\boldsymbol{s}|^q}{q} + \frac{|\boldsymbol{d}|^p}{p}$$

which not only verifies, but also motivates the assumption (3.10) *above. To verify that the other models fulfill* (3.10)*, we refer to [20, Lemma 1.1] and [13, Lemma B.1]. Finally, the responses given by*

$$\begin{aligned}\boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{s} - \frac{\boldsymbol{d}}{\left(1+|\boldsymbol{d}|^a\right)^{\frac{1}{a}}}, \Big\}\ \textit{flux limiting p-Laplace}\\ \boldsymbol{h}(\boldsymbol{s},\boldsymbol{d}) &= \boldsymbol{d} - \frac{\boldsymbol{s}}{\left(1+|\boldsymbol{s}|^b\right)^{\frac{1}{b}}}, \Big\}\ \textit{gradient limiting p-Laplace}\end{aligned}$$

with some $a, b \in (0,\infty)$ give automatically $\boldsymbol{\sigma} \in [L^\infty(\Omega)]^d$, $\nabla u \in [L^1(\Omega)]^d$, respectively $\boldsymbol{\sigma} \in [L^1(\Omega)]^d$, $\nabla u \in [L^\infty(\Omega)]^d$ and concern the limit cases $p = 1$, $p = \infty$. We refer to [13], where such models are summarized in the context of fluid mechanics, and [22, 41] for examples from solid mechanics. This general setting with implicit function $\boldsymbol{h}$ is, for example, interesting to employ mixed finite element methods. In fluid mechanics context, this has been studied in [38, 14].

The above rather complex example still fits perfectly into our setting. Indeed, let $\boldsymbol{\sigma}_h \in [L^q(\Omega)]^d$ be an arbitrary approximation to $\boldsymbol{\sigma}$. Then we can define a linear functional $\mathcal{R}$ on the space V as

$$\langle \mathcal{R}, v\rangle_{V',V} := (f,v) - (\boldsymbol{\sigma}_h, \nabla v) \qquad v \in V.$$

Note that the Hölder inequality and the Poincaré–Friedrichs inequality (2.5)*, used in the entire domain Ω on the space V, imply that*

$$|\langle \mathcal{R}, v\rangle| \leq (\|f\|_q C_{\mathrm{PF},p,\Omega} h_\Omega + \|\boldsymbol{\sigma}_h\|_q)\|\nabla v\|_p.$$

Consequently, $\mathcal{R}$ is indeed bounded, $\mathcal{R} \in V'$. To complement, let also $u_h \in W^{1,p}(\Omega)$, $u_h - u^{\mathrm{D}} \in V$, be an arbitrary approximation to u. Then one in general also wishes to measure a deviation from equality (3.9c) *when $\boldsymbol{\sigma}_h$ together with u_h are plugged therein in place of $\boldsymbol{\sigma}$ and u. There are various ways to evaluate this error; compare, e.g., [40, 14].*

For the rest of this example, we limit ourselves to the following specific but important subcase: $1 < p < \infty$ *and the implicit function* $\boldsymbol{h}$ *admits an explicit continuous representation* $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{d})$;[3] *more precisely we assume that all solutions* $(\boldsymbol{s}, \boldsymbol{d}) \in \mathbb{R}^d \times \mathbb{R}^d$ *of* $\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{0}$ *are given by* $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{d})$ *with continuous* $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^d$. *Then the weak formulation of problem* (3.9) *simplifies to: find* $u \in W^{1,p}(\Omega)$ *such that*

$$u - u^{\mathrm{D}} \in V, \tag{3.11a}$$

$$(\boldsymbol{\sigma}(\nabla u), \nabla v) = (f, v) \qquad \forall v \in V \tag{3.11b}$$

and admits at least one weak solution under classical assumptions.[4] *This gives rise to the standard notion of the residual* $\mathcal{R}$ *of an arbitrary function* $u_h \in W^{1,p}(\Omega)$ *such that* $u_h - u^{\mathrm{D}} \in V$, *defined via*

$$\langle \mathcal{R}, v \rangle_{V',V} := (f, v) - (\boldsymbol{\sigma}(\nabla u_h), \nabla v) \qquad v \in V. \tag{3.12}$$

The Hölder inequality and (2.5) *again imply that* $\mathcal{R} \in V'$, *since*

$$|\langle \mathcal{R}, v \rangle| \le (\|f\|_q C_{\mathrm{PF},p,\Omega} h_\Omega + \|\boldsymbol{\sigma}(\nabla u_h)\|_q) \|\nabla v\|_p.$$

Here, actually, $\mathcal{R} = 0$ *if and only if* u_h *solves* (3.11b). *Then* $\|\mathcal{R}\|_{V'}$ *is the* intrinsic distance of u_h to u, *the* dual norm of the residual. *Remark that this problem can also be cast in the form of Example 3.1, taking* $\boldsymbol{\xi} := \boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)$, *with any* $u \in W^{1,p}(\Omega)$ *solving* (3.11).

3.3 Motivation

We now give four remarks motivating our main question whether $\|\mathcal{R}\|_{V'}$, a priori just a number defined for any $\mathcal{R} \in V'$, expressing its size over the entire computational domain Ω, can be bounded from above and from below by the sizes $\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}$ of $\mathcal{R}$ localized over the patches $\omega_{\boldsymbol{a}}$.

Remark 3.3 (Localization of the $\boldsymbol{L}^q(\Omega)$-norm error in the fluxes)**.** *Consider* $\mathcal{R}$ *given by* (3.12) *from Example 3.2 in the finite element context of Remark 2.1. We immediately obtain from* (3.6a) *and* (3.6b)

$$\|\mathcal{R}\|_{V'} \le \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q, \tag{3.13a}$$

$$\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \le \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_{q,\omega_{\boldsymbol{a}}} \qquad \forall \boldsymbol{a} \in \mathcal{V}_h, \tag{3.13b}$$

and observe that the flux error norm on the right-hand side of (3.13a) *localizes, as in* (2.7), *into the right-hand sides of* (3.13b) *by the formula*

$$\|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q = \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_{q,\omega_{\boldsymbol{a}}}^q \right\}^{\frac{1}{q}}. \tag{3.14}$$

Note that, unfortunately, it is unclear when (3.14) *is, up to a constant, bounded back by* $\|\mathcal{R}\|_{V'}$, *so that these considerations do not give an answer to the question of localization of* $\|\mathcal{R}\|_{V'}$.

Remark 3.4 ($W_0^{1,p}(\Omega)$-norm error localization)**.** *Remark that similarly to* (1.2), *there always holds, for* $1 \le p < \infty$,

$$\|\nabla v\|_p = \left\{ \sum_{K \in \mathcal{T}_h} \|\nabla v\|_{p,K}^p \right\}^{\frac{1}{p}}, \qquad v \in V.$$

[3] If $\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{d}) = \boldsymbol{0}$ does not admit explicit solution $\boldsymbol{s} = \boldsymbol{\sigma}(\boldsymbol{d})$, which happens for some examples given above, one can approximate up to (in certain sense) arbitrary precision, by explicit relation $\boldsymbol{s} = \boldsymbol{\sigma}^\epsilon(\boldsymbol{d})$, and later pass in the limit $\epsilon \to 0+$. This is an approach of many studies, ranging from PDE analysis to a priori convergence of finite element schemes; see, e.g., [36, 19, 20, 30, 40].

[4] This holds, for example, if

1. $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathbb{R}^d$ is continuous,
2. $\boldsymbol{\sigma}(\boldsymbol{0}) = \boldsymbol{0}$,
3. $(\boldsymbol{\sigma}(\boldsymbol{d}_1) - \boldsymbol{\sigma}(\boldsymbol{d}_2)) \cdot (\boldsymbol{d}_1 - \boldsymbol{d}_2) \ge 0$ for all $\boldsymbol{d}_1$, $\boldsymbol{d}_2 \in \mathbb{R}^d$,
4. $C_1 |\boldsymbol{d}|^p \le \boldsymbol{\sigma}(\boldsymbol{d}) \cdot \boldsymbol{d}$, for all $\boldsymbol{d} \in \mathbb{R}^d$,
5. $|\boldsymbol{\sigma}(\boldsymbol{d})| \le C_2 (1 + |\boldsymbol{d}|)^{p-1}$ for all $\boldsymbol{d} \in \mathbb{R}^d$.

See, e.g. [42].

In particular, in the context of Example 3.2, on meshes from Remark 2.1, for $1 < p < \infty$,

$$\|\nabla(u - u_h)\|_p = \left\{ \sum_{K \in \mathcal{T}_h} \|\nabla(u - u_h)\|_{p,K}^p \right\}^{\frac{1}{p}}. \tag{3.15}$$

The $W_0^{1,p}(\Omega)$*-norm* $\|\nabla \cdot\|_p$ *is always localizable, but it seems difficult/suboptimal to derive a posteriori error estimates of the form* (1.8), (1.10) *for* $\|\nabla(u - u_h)\|_p$ *in place of* $\|\mathcal{R}\|_{V'}$, *see, e.g., the discussions in Belenki* et al. *[10] and [33].*

Remark 3.5 (Energy difference/quasi-norm error localization)**.** *As mentioned in the introduction, still in the context* (3.11) *of Example 3.2, there are other possible substitutes used in both a priori and a posteriori error analysis. Besides* $W_0^{1,p}(\Omega)$*-norm error* $\|\nabla(u - u_h)\|_p$, *the energy difference* $\mathcal{E}(u_h) - \mathcal{E}(u)$, *where the energy is defined by* (1.7), *is used mostly for the problem involving the p-Laplace or its nondegenerate/nonsingular modifications. Following Kreuzer [29, Lemma 16] or Belenki* et al. *[10, Lemma 3.2], there holds*

$$\mathcal{E}(u_h) - \mathcal{E}(u) \approx \|\nabla(u - u_h)\|_{(p)}^2 \approx \|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\|^2, \tag{3.16}$$

where $\|\cdot\|_{(p)}$ *is the quasi-norm of Barrett and Liu [6, 7] and* $\boldsymbol{F}(\boldsymbol{v}) := |\boldsymbol{v}|^{\frac{p-2}{2}}\boldsymbol{v}$. *Here* $\|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\|^2 = \sum_{K \in \mathcal{T}_h} \|\boldsymbol{F}(\nabla u) - \boldsymbol{F}(\nabla u_h)\|_K^2$ *localizes immediately. However, unfortunately, the constants hidden in* (3.16) *depend on the Lebesgue exponent p.*

Remark 3.6 (Localization of the p-Laplacian lifting of $\mathcal{R}$)**.** *Let* $1 < p < \infty$. *Let* $\mathscr{r} \in V$ *be the analogue of the Riesz representation of the functional* $\mathcal{R}$ *by the p-Laplacian solve on* Ω, *i.e.,* $\mathscr{r} \in V$ *is such that*

$$(|\nabla \mathscr{r}|^{p-2}\nabla \mathscr{r}, \nabla v) = \langle \mathcal{R}, v\rangle_{V',V} \qquad \forall v \in V. \tag{3.17}$$

Then, we readily obtain

$$\|\nabla \mathscr{r}\|_p^p = (|\nabla \mathscr{r}|^{p-2}\nabla \mathscr{r}, \nabla \mathscr{r}) = \langle \mathcal{R}, \mathscr{r}\rangle_{V',V} = \|\mathcal{R}\|_{V'}^q. \tag{3.18}$$

Consequently, on meshes $\mathcal{T}_h$ *from Remark 2.1,*

$$\|\nabla \mathscr{r}\|_p = \left\{ \sum_{K \in \mathcal{T}_h} \|\nabla \mathscr{r}\|_{p,K}^p \right\}^{\frac{1}{p}} \tag{3.19}$$

suggests itself as a way to measure the error with localization and a posteriori estimate of the form (1.8). *Also an equivalent of* (1.10),

$$\eta_K(u_h) \le C \left\{ \sum_{K' \in \mathcal{T}_K} \|\nabla \mathscr{r}\|_{K'}^p \right\}^{\frac{1}{q}},$$

would hold. The trouble here is that (3.17) *is a* nonlocal *problem, obtained itself by a global solve. Remark also that the definition of the lifting* $\mathscr{r}$ *by* (3.17) *is dictated by the choice of the space V in* (3.1) *together with its norm* $\|\nabla \cdot\|_p$.[5]

3.4 Main result

Recall that $1 \le p \le \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, $V = W_0^{1,p}(\Omega)$, the partition $\cup_{\boldsymbol{a} \in \mathcal{V}_h} \omega_{\boldsymbol{a}}$ covers the domain Ω with maximal overlap N_{ov}, the patches $\omega_{\boldsymbol{a}}$ are indexed by the vertices $\boldsymbol{a}$ where $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$ lies inside Ω and $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}$ on the boundary of Ω, and that the constant $C_{\mathrm{cont,PF}}$ from (2.6) is supposed uniformly bounded for different partitions.

Our localization result is:

[5] Consider an alternative choice of the space, $V := \{v \in W^{1,p}(\Omega); (v, 1) = 0\}$ with the norm $\|\nabla v\|_p$. We have $\|\mathcal{R}\|_{V'} := \sup_{v \in V; \|\nabla v\|_p = 1} \langle \mathcal{R}, v\rangle_{V',V}$, as in (3.2). For $\mathcal{R} \in V'$, one can define a lifting $\mathscr{r} \in V$ as a solution of the Neumann p-Laplace problem $-\mathrm{div}(|\nabla \mathscr{r}|^{p-2}\nabla \mathscr{r}) = \mathcal{R}$ in Ω; $|\nabla \mathscr{r}|^{p-2}\nabla \mathscr{r}\boldsymbol{n} = 0$ on $\partial\Omega$; $(\mathscr{r}, 1) = 0$. The weak formulation (3.17), the $W^{1,p}(\Omega)$-norm equality (3.18), and the localization (3.19) hold with the appropriate replacement of V.

Theorem 3.7 (Localization of dual norms of functionals with $\psi_{\boldsymbol{a}}$-orthogonality)**.** *Let $\mathcal{R} \in V'$ be arbitrary. Let*

$$\langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{V',V} = 0 \qquad \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}. \tag{3.20}$$

Then, when $1 < p \leq \infty$,

$$\|\mathcal{R}\|_{V'} \leq N_{\mathrm{ov}} C_{\mathrm{cont,PF}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \right\}^{\frac{1}{q}}, \tag{3.21a}$$

$$\left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \right\}^{\frac{1}{q}} \leq \|\mathcal{R}\|_{V'}, \tag{3.21b}$$

and, when $p = 1$,

$$\|\mathcal{R}\|_{V'} \leq N_{\mathrm{ov}} C_{\mathrm{cont,PF}} \max_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}, \tag{3.22a}$$

$$\max_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \leq \|\mathcal{R}\|_{V'}. \tag{3.22b}$$

Condition (3.20) *is actually only needed in* (3.21a) *and* (3.22a).

Proof. We first show (3.21a) and (3.22a). Let $v \in V$ with $\|\nabla v\|_p = 1$ be fixed. The partition of unity (2.2), the linearity of $\mathcal{R}$, definition (3.3), and the orthogonality requirement (3.20) give

$$\begin{aligned} \langle \mathcal{R}, v \rangle_{V',V} &= \sum_{\boldsymbol{a} \in \mathcal{V}_h} \langle \mathcal{R}, \psi_{\boldsymbol{a}} v \rangle_{V',V} = \sum_{\boldsymbol{a} \in \mathcal{V}_h} \langle \mathcal{R}, \psi_{\boldsymbol{a}} v \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} \\ &= \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}} \langle \mathcal{R}, \psi_{\boldsymbol{a}} (v - \Pi_{0,\omega_{\boldsymbol{a}}} v) \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} + \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}} \langle \mathcal{R}, \psi_{\boldsymbol{a}} v \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}, \end{aligned} \tag{3.23}$$

where $\Pi_{0,\omega_{\boldsymbol{a}}} v$ is the mean value of the test function v on the patch $\omega_{\boldsymbol{a}}$. There holds $(v - \Pi_{0,\omega_{\boldsymbol{a}}} v)|_{\omega_{\boldsymbol{a}}} \in W^{1,p}_*(\omega_{\boldsymbol{a}})$, where $W^{1,p}_*(\omega_{\boldsymbol{a}})$ is defined by (2.4), and $(\psi_{\boldsymbol{a}}(v - \Pi_{0,\omega_{\boldsymbol{a}}} v))|_{\omega_{\boldsymbol{a}}} \in V^{\boldsymbol{a}}$ for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$. Thus, using (3.4) and Lemma 2.4 yields, for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$,

$$\begin{aligned} \langle \mathcal{R}, \psi_{\boldsymbol{a}} (v - \Pi_{0,\omega_{\boldsymbol{a}}} v) \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} &\leq \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \|\nabla(\psi_{\boldsymbol{a}} (v - \Pi_{0,\omega_{\boldsymbol{a}}} v))\|_{p,\omega_{\boldsymbol{a}}} \\ &\leq C_{\mathrm{cont,PF}} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \|\nabla(v - \Pi_{0,\omega_{\boldsymbol{a}}} v)\|_{p,\omega_{\boldsymbol{a}}} \\ &= C_{\mathrm{cont,PF}} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \|\nabla v\|_{p,\omega_{\boldsymbol{a}}}. \end{aligned}$$

For $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}$, there holds $v|_{\omega_{\boldsymbol{a}}} \in W^{1,p}_*(\omega_{\boldsymbol{a}})$ and $(\psi_{\boldsymbol{a}} v)|_{\omega_{\boldsymbol{a}}} \in V^{\boldsymbol{a}}$. Hence, similarly, we obtain

$$\langle \mathcal{R}, \psi_{\boldsymbol{a}} v \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF}} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \|\nabla v\|_{p,\omega_{\boldsymbol{a}}}.$$

Thus, the Hölder inequality gives, for $1 < p < \infty$,

$$\langle \mathcal{R}, v \rangle_{V',V} \leq N_{\mathrm{ov}} C_{\mathrm{cont,PF}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \right\}^{\frac{1}{q}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\nabla v\|^p_{p,\omega_{\boldsymbol{a}}} \right\}^{\frac{1}{p}}.$$

Combining (2.3) used for ∇v with (3.2) now implies the result if $1 < p < \infty$. Cases $p = 1$ and $p = \infty$ are obvious modifications.

We now pass to (3.21b) and (3.22b). First assume that $1 < p \leq \infty$. From (3.4) we deduce that for any $\boldsymbol{a} \in \mathcal{V}_h$

$$\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} = \sup_{v \in V^{\boldsymbol{a}};\, \|\nabla v\|_{p,\omega_{\boldsymbol{a}}} = \|\mathcal{R}\|^{q-1}_{(V^{\boldsymbol{a}})'}} \langle \mathcal{R}, v \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}.$$

For a fixed $\boldsymbol{a} \in \mathcal{V}_h$, we can characterize the supremum by a sequence $\{v_j^{\boldsymbol{a}}\}_{j=1}^{\infty} \subset V^{\boldsymbol{a}}$ with

$$\|\nabla v_j^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}} = \|\mathcal{R}\|^{q-1}_{(V^{\boldsymbol{a}})'} \qquad \text{(with convention } 0^0 = 1) \tag{3.24}$$

and

$$\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} = \lim_{j\to\infty} \langle \mathcal{R}, \mathscr{r}^{\boldsymbol{a}}_j \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}. \tag{3.25}$$

After summing over $\boldsymbol{a} \in \mathcal{V}_h$, dividing by N_{ov}, and using (3.3) together with the linearity of $\mathcal{R}$, we can estimate

$$\frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} = \lim_{j\to\infty} \left\langle \mathcal{R}, \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \mathscr{r}^{\boldsymbol{a}}_j \right\rangle_{V',V} \le \lim_{j\to\infty} \|\mathcal{R}\|_{V'} \left\| \frac{1}{N_{\mathrm{ov}}} \nabla \sum_{\boldsymbol{a}\in\mathcal{V}_h} \mathscr{r}^{\boldsymbol{a}}_j \right\|_p .$$

Using Lemma 2.5 and (3.24) we get

$$\frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \le \begin{cases} \|\mathcal{R}\|_{V'} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \right\}^{\frac{1}{p}} & 1 < p < \infty, \\ \|\mathcal{R}\|_{V'} & p = \infty, \end{cases}$$

which proves (3.21b). The case (3.22b) follows easily by (3.2)–(3.4). □

3.5 Remarks

We collect here a couple of remarks associated with Theorem 3.7.

Remark 3.8 (Expressing local norms using p-Laplace local liftings)**.** *Let $1 < p < \infty$. In this case $V^{\boldsymbol{a}}$ is reflexive, so for the sequence $\{\mathscr{r}^{\boldsymbol{a}}_j\}_{j=1}^\infty$ from the proof of Theorem 3.7, there is a subsequence which converges weakly to some $\mathscr{r}^{\boldsymbol{a}} \in V^{\boldsymbol{a}}$ with $\|\nabla \mathscr{r}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}} \le \liminf_{j\to\infty} \|\nabla \mathscr{r}^{\boldsymbol{a}}_j\|_{p,\omega_{\boldsymbol{a}}} = \|\mathcal{R}\|^{q-1}_{(V^{\boldsymbol{a}})'}$ thanks to weak lower semicontinuity of norm and* (3.24)*. On the other hand, from* (3.25) *and the weak convergence, we conclude that*

$$\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} = \langle \mathcal{R}, \mathscr{r}^{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}, \tag{3.26}$$

which implies that $\|\mathcal{R}\|^{q-1}_{(V^{\boldsymbol{a}})'} \le \|\nabla \mathscr{r}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}$. Hence, altogether we have that

$$\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} = \|\nabla \mathscr{r}^{\boldsymbol{a}}\|^p_{p,\omega_{\boldsymbol{a}}}. \tag{3.27}$$

Moreover, as $V^{\boldsymbol{a}}$ (or, equivalently, $\|\nabla\cdot\|^p_{p,\omega_{\boldsymbol{a}}}$) is a strictly convex (in fact uniformly convex) space, when $1 < p < \infty$, $\mathscr{r}^{\boldsymbol{a}} \in V^{\boldsymbol{a}}$ with properties (3.26), (3.27) *is unique. For proof assume that $\mathcal{R} \ne 0$ on $V^{\boldsymbol{a}}$ (the case $\mathcal{R} = 0$ on $V^{\boldsymbol{a}}$ is trivial) and that there is $\mathscr{s}^{\boldsymbol{a}} \ne \alpha \mathscr{r}^{\boldsymbol{a}}$ and $\mathscr{s}^{\boldsymbol{a}}$ satisfies* (3.26) *and* (3.27) *with $\mathscr{s}^{\boldsymbol{a}}$ in place of $\mathscr{r}^{\boldsymbol{a}}$. Define $\mathscr{z}^{\boldsymbol{a}} := \frac{\mathscr{r}^{\boldsymbol{a}} + \mathscr{s}^{\boldsymbol{a}}}{\|\nabla(\mathscr{r}^{\boldsymbol{a}} + \mathscr{s}^{\boldsymbol{a}})\|_{p,\omega_{\boldsymbol{a}}}} \in V^{\boldsymbol{a}}$ with $\|\nabla \mathscr{z}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}} = 1$ and observe using* (3.26), (3.27) *and the strict convexity $\|\nabla(\mathscr{r}^{\boldsymbol{a}} + \mathscr{s}^{\boldsymbol{a}})\|_{p,\omega_{\boldsymbol{a}}} < \|\nabla \mathscr{r}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}} + \|\nabla \mathscr{s}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}$ that $\langle \mathcal{R}, \mathscr{z}^{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} > \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}$, which is a contradiction with* (3.4).

It is easy to check that the unique solution $\mathscr{r}^{\boldsymbol{a}} \in V^{\boldsymbol{a}}$ of (3.26), (3.27) *is in fact the solution of p-Laplacian solve on the patch $\omega_{\boldsymbol{a}}$:*

$$(|\nabla \mathscr{r}^{\boldsymbol{a}}|^{p-2} \nabla \mathscr{r}^{\boldsymbol{a}}, \nabla v)_{\omega_{\boldsymbol{a}}} = \langle \mathcal{R}, v \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} \qquad \forall v \in V^{\boldsymbol{a}}. \tag{3.28}$$

Note that the above reasoning about the existence and uniqueness of representation (3.28), *which in its generality referred only to reflexivity and strict convexity of $V^{\boldsymbol{a}}$, applies also to global representation of $\mathcal{R}$ on V, as defined by* (3.17)*; see also footnote 5 on page 59.*

Remark 3.9 (Localization based on weighted Poincaré–Friedrichs inequalities)**.** *Poincaré–Friedrichs inequalities can be derived for the weighted $L^p(\Omega)$-norm of v on $\omega_{\boldsymbol{a}}$, $\|\psi_{\boldsymbol{a}}^{\frac{1}{p}} v\|_{p,\omega_{\boldsymbol{a}}}$ in place of $\|v\|_{p,\omega_{\boldsymbol{a}}}$ in* (2.5)*, see Chua and Wheeden [26] and Veeser and Verfürth [51]. Then, in the spirit of Carstensen and Funken [23] and Veeser and Verfürth [50], weighted equivalents of Lemma 2.4 and Theorem 3.7 could be given. This might reduce the size of the constants in* (3.21)–(3.22)*, originating from overlapping of the supports of the test functions $\psi_{\boldsymbol{a}}$, at the price of making the formulas a little more involved.*

We finally show that inequality (3.21b) can be split into local contributions when passing from dual norms of the functional $\mathcal{R}$ to its liftings.

Remark 3.10 (Splitting (3.21b) into local contributions using lifted norms)**.** *Let* $1 < p < \infty$ *and let* $\mathcal{R} \in V'$ *and* $\boldsymbol{a} \in \mathcal{V}_h$ *be given. Define the* global lifting $\mathscr{r} \in V$ *of the functional* $\mathcal{R}$ *by* (3.17) *and the* local lifting $\mathscr{r}^{\boldsymbol{a}} \in V^{\boldsymbol{a}}$ *by* (3.28)*. Then it holds*

$$\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} = \|\nabla \mathscr{r}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}^{p-1} \le \|\nabla \mathscr{r}\|_{p,\omega_{\boldsymbol{a}}}^{p-1}. \tag{3.29}$$

Indeed, the equality has been shown in equation (3.27) *and the inequality follows using definition* (3.4)*, definition of the global lifting* (3.17)*, and the Hölder inequality*

$$\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} = \sup_{v \in V^{\boldsymbol{a}};\, \|\nabla v\|_{p,\omega_{\boldsymbol{a}}}=1} \langle \mathcal{R}, v\rangle_{V',V} = \sup_{v \in V^{\boldsymbol{a}};\, \|\nabla v\|_{p,\omega_{\boldsymbol{a}}}=1} (|\nabla \mathscr{r}|^{p-2}\nabla \mathscr{r}, \nabla v)_{\omega_{\boldsymbol{a}}} \le \|\nabla \mathscr{r}\|_{p,\omega_{\boldsymbol{a}}}^{p-1}.$$

Note that summing (3.29) *in q-th power over all vertices* $\mathcal{V}_h$ *and using* (2.3a) *and* (3.18) *one gets* (3.21b) *as a trivial consequence.*

4 Extensions

This section collects various extensions of the main result of Theorem 3.7.

4.1 Localization without the orthogonality condition

We begin by a simple generalization of Theorem 3.7 to the case without orthogonality (3.20) to the partition of unity functions $\psi_{\boldsymbol{a}}$.

Theorem 4.1 (Simple localization of dual norms of functionals without $\psi_{\boldsymbol{a}}$-orthogonality)**.** *Let* $\mathcal{R} \in V'$ *be arbitrary and define*

$$r^{\boldsymbol{a}} := \frac{h_\Omega C_{\mathrm{PF},p,\Omega}}{|\omega_{\boldsymbol{a}}|^{\frac{1}{p}}} |\langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}}|. \tag{4.1}$$

Then, when $1 < p \le \infty$*,*

$$\|\mathcal{R}\|_{V'} \le N_{\mathrm{ov}} C_{\mathrm{cont,PF}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}^q \right\}^{\frac{1}{q}} + N_{\mathrm{ov}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}} (r^{\boldsymbol{a}})^q \right\}^{\frac{1}{q}}, \tag{4.2a}$$

$$\left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}^q \right\}^{\frac{1}{q}} \le \|\mathcal{R}\|_{V'}, \tag{4.2b}$$

and, when $p = 1$*,*

$$\|\mathcal{R}\|_{V'} \le N_{\mathrm{ov}} C_{\mathrm{cont,PF}} \max_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} + N_{\mathrm{ov}} \max_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}} r^{\boldsymbol{a}}, \tag{4.3a}$$

$$\max_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \le \|\mathcal{R}\|_{V'}. \tag{4.3b}$$

Proof. Estimates (4.2b) and (4.3b) have been proven in Theorem 3.7. Estimates (4.2a) and (4.3a) are proven along the lines of Theorem 3.7, counting for the additional nonzero term

$$\sum_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}} (\Pi_{0,\omega_{\boldsymbol{a}}} v) \langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}} \tag{4.4}$$

in (3.23). For each $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$, the Hölder inequality gives

$$|\omega_{\boldsymbol{a}}|^{\frac{1}{p}} (\Pi_{0,\omega_{\boldsymbol{a}}} v) = |\omega_{\boldsymbol{a}}|^{\frac{1}{p}} (v, 1)_{\omega_{\boldsymbol{a}}} |\omega_{\boldsymbol{a}}|^{-1} \le |\omega_{\boldsymbol{a}}|^{\frac{1}{p}} \|v\|_{p,\omega_{\boldsymbol{a}}} |\omega_{\boldsymbol{a}}|^{\frac{1}{q}} |\omega_{\boldsymbol{a}}|^{-1} = \|v\|_{p,\omega_{\boldsymbol{a}}}.$$

Thus, the Hölder inequality, the Poincaré–Friedrichs inequality (2.5) used in the entire domain Ω on the space V, and (2.3) lead to, for $1 < p < \infty$,

$$
\begin{aligned}
\sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} (\Pi_{0,\omega_{\boldsymbol{a}}} v)\langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}} &= \sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} |\omega_{\boldsymbol{a}}|^{-\frac{1}{p}} \langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}} |\omega_{\boldsymbol{a}}|^{\frac{1}{p}} (\Pi_{0,\omega_{\boldsymbol{a}}} v) \\
\leq N_{\mathrm{ov}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} (|\omega_{\boldsymbol{a}}|^{-\frac{1}{p}} |\langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}}|)^q \right\}^{\frac{1}{q}} &\left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} \|v\|_{p,\omega_{\boldsymbol{a}}}^p \right\}^{\frac{1}{p}} \\
\leq N_{\mathrm{ov}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} (r^{\boldsymbol{a}})^q \right\}^{\frac{1}{q}} \|\nabla v\|_p,&
\end{aligned}
$$

and (3.2) gives the assertion. Cases $p = 1$ and $p = \infty$ are proved with obvious modifications. □

This result implies the following remark:

Remark 4.2 (h-unstable localization of dual norms of functionals)**.** *Observe that in* (4.2a) *and* (4.3a)*, we can apply* $|\langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}}| \leq \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \|\nabla\psi_{\boldsymbol{a}}\|_p$ *and the Hölder inequality in order to arrive at*

$$
\|\mathcal{R}\|_{V'} \leq N_{\mathrm{ov}} C_h^{\overset{\ddagger}{\infty}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}^q \right\}^{\frac{1}{q}}, \qquad 1 < p \leq \infty, \tag{4.5}
$$

$$
\|\mathcal{R}\|_{V'} \leq N_{\mathrm{ov}} C_h^{\overset{\ddagger}{\infty}} \max_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}, \qquad p = 1, \tag{4.6}
$$

with

$$
C_h^{\overset{\ddagger}{\infty}} := \Big(C_{\mathrm{cont,PF}} + h_\Omega C_{\mathrm{PF},p,\Omega} \max_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}} \Big). \tag{4.7}
$$

Whereas h_Ω *and* $C_{\mathrm{PF},p,\Omega}$ *do not depend on the partition and* $C_{\mathrm{cont,PF}}$ *is uniformly bounded for regular partitions, there typically holds* $\max_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}} \approx h^{-1}$*, so that* $C_h^{\overset{\ddagger}{\infty}}$ *explodes for small patches* $\omega_{\boldsymbol{a}}$*,* $\boldsymbol{a} \in \mathcal{V}_h$*. We note that one can actually estimate a little more sharply with* $C_h^{\overset{\ddagger}{\infty}} = 1 + h_\Omega C_{\mathrm{PF},p,\Omega} \max_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}$*.*

Estimates (4.2a) and (4.3a) of Theorem 4.1 take a simple form but, unfortunately, as Example 4.6 below shows, the second term in (4.2a) may severely overestimate $\|\mathcal{R}\|_{V'}$. Correspondingly, (4.5) and (4.6) of Remark 4.2 blow up with mesh refinement due to presence of $\max_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}$ in (4.7). We discuss in Example 4.6 that it is related to l^2-norm estimates of the algebraic residual vector from numerical linear algebra; in both cases, the local contributions are first taken in *absolute value* in (4.1) and then the size of the *resulting algebraic vector* is measured in the second term in (4.2a). The following estimate, obtained while employing the ideas of [39, Section 7.3] and [44], removes this deficiency, while *first summing* the local contributions and then constructing a *discrete* $\boldsymbol{H}^q(\mathrm{div},\Omega)$*-lifting.*

Theorem 4.3 (Improved localization of dual norms of functionals without $\psi_{\boldsymbol{a}}$-orthogonality)**.** *Let* $\mathcal{R} \in V'$ *be arbitrary and define* $r_h \in \mathbb{P}_0(\mathcal{T}_h)$ *to be the piecewise constant function with respect to the partition* $\mathcal{T}_h$ *given by*

$$
r_h|_K := \sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}\cap\mathcal{V}_K} \frac{1}{|\omega_{\boldsymbol{a}}|} \langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}} \qquad \forall K \in \mathcal{T}_h. \tag{4.8}
$$

Let $\boldsymbol{\sigma}_{h,\mathrm{alg}} \in \boldsymbol{H}^q(\mathrm{div},\Omega) := \{\boldsymbol{v} \in [L^q(\Omega)]^d; \mathrm{div}\, \boldsymbol{v} \in L^q(\Omega)\}$ *be arbitrary but such that*

$$
\mathrm{div}\, \boldsymbol{\sigma}_{h,\mathrm{alg}} = r_h. \tag{4.9}
$$

Then, when $1 < p \leq \infty$,

$$\|\mathcal{R}\|_{V'} \leq N_{\rm ov} C_{\rm cont,PF} \left\{ \frac{1}{N_{\rm ov}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \right\}^{\frac{1}{q}} + \|\boldsymbol{\sigma}_{h,\rm alg}\|_q, \tag{4.10a}$$

$$\left\{ \frac{1}{N_{\rm ov}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'} \right\}^{\frac{1}{q}} \leq \|\mathcal{R}\|_{V'}, \tag{4.10b}$$

and, when $p = 1$,

$$\|\mathcal{R}\|_{V'} \leq N_{\rm ov} C_{\rm cont,PF} \max_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} + \|\boldsymbol{\sigma}_{h,\rm alg}\|_\infty, \tag{4.11a}$$

$$\max_{\boldsymbol{a} \in \mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \leq \|\mathcal{R}\|_{V'}. \tag{4.11b}$$

Proof. The proof consists in finding an alternative, sharper bound on the term (4.4) above. Let $v \in V$ with $\|\nabla v\|_p = 1$ be fixed. Note that, for each interior vertex $\boldsymbol{a} \in \mathcal{V}_h^{\rm int}$,

$$(\Pi_{0,\omega_{\boldsymbol{a}}} v) \langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} = \frac{1}{|\omega_{\boldsymbol{a}}|} (\langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}, v)_{\omega_{\boldsymbol{a}}}.$$

Hence, considering $\frac{1}{|\omega_{\boldsymbol{a}}|} \langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}$ as constant on $\omega_{\boldsymbol{a}}$ and zero elsewhere and using definition (4.8),

$$\begin{aligned} \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\rm int}} (\Pi_{0,\omega_{\boldsymbol{a}}} v) \langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} &= \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\rm int}} \left(\frac{1}{|\omega_{\boldsymbol{a}}|} \langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}}, v \right) = (r_h, v) \\ &= (\operatorname{div} \boldsymbol{\sigma}_{h,\rm alg}, v) = -(\boldsymbol{\sigma}_{h,\rm alg}, \nabla v) \leq \|\boldsymbol{\sigma}_{h,\rm alg}\|_q \|\nabla v\|_p = \|\boldsymbol{\sigma}_{h,\rm alg}\|_q, \end{aligned}$$

where we have also applied the requirement (4.9), the Green theorem, and the Hölder inequality. Actually, generalizing [39, Theorem 5.5] to the present setting, it follows that, at least for $1 < p < \infty$,

$$\sup_{v \in V;\, \|\nabla v\|_p = 1} \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\rm int}} (\Pi_{0,\omega_{\boldsymbol{a}}} v) \langle \mathcal{R}, \psi_{\boldsymbol{a}} \rangle_{(V^{\boldsymbol{a}})', V^{\boldsymbol{a}}} = \min_{\boldsymbol{\sigma}_{h,\rm alg} \in \boldsymbol{H}^q({\rm div},\Omega);\, \operatorname{div} \boldsymbol{\sigma}_{h,\rm alg} = r_h} \|\boldsymbol{\sigma}_{h,\rm alg}\|_q,$$

so that this estimate is as sharp as possible.

□

Example 4.4 (Construction of $\boldsymbol{\sigma}_{h,\rm alg}$)**.** *Several practical constructions of* $\boldsymbol{\sigma}_{h,\rm alg}$ *in finite-dimensional subspaces of* $\boldsymbol{H}^q({\rm div},\Omega)$ *in the context of simplicial or parallelepipedal meshes of Remark 2.1 are possible, employing the lowest-order Raviart–Thomas–Nédélec (*$\mathbf{RTN}_0$*) space, cf. [17] and the references therein. A construction with a cost linear in terms of the number of mesh elements of* $\mathcal{T}_h$ *has been proposed in [39, Section 7.3]. It consists in a (sequential) sweep through all mesh elements in a proper order. Numerically often much sharper construction has been proposed in [44, Definition 6.3]. It needs a hierarchy of meshes of whose* $\mathcal{T}_h$ *is a refinement, in the multigrid spirit, and consists in an exact solve on the coarsest mesh and a (parallel) sweep through all mesh vertices on all mesh levels. This latter construction can be shown to be an optimal estimate (giving both upper and lower (up to a constant) bounds)(work in progress). Note that although references [39, 44] consider the Hilbertian setting* $p = 2$*, there is no structural loss in passing to* $p \neq 2$*, see [31, 33] and the references therein.*

Remark 4.5 (Localization of $\|\boldsymbol{\sigma}_{h,\rm alg}\|_q$)**.** *Note that from* (2.3)*, one has* $\|\boldsymbol{\sigma}_{h,\rm alg}\|_q^q \leq \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\boldsymbol{\sigma}_{h,\rm alg}\|^q_{q,\omega_{\boldsymbol{a}}} \leq N_{\rm ov} \|\boldsymbol{\sigma}_{h,\rm alg}\|_q^q$*. In the context of Remark 2.1, actually* $\|\boldsymbol{\sigma}_{h,\rm alg}\|_q = \left\{ \frac{1}{N_{\rm ov}} \sum_{\boldsymbol{a} \in \mathcal{V}_h} \|\boldsymbol{\sigma}_{h,\rm alg}\|^q_{q,\omega_{\boldsymbol{a}}} \right\}^{\frac{1}{q}}$*. Thus, the second terms on the right-hand sides of* (4.10a) *and* (4.11a) *are fully localizable.*

Example 4.6 (Link of estimate of Theorem 4.1 to the l^2-norm of the algebraic residual vector when $p = 2$ and their deficiency[6])**.** *Consider* $p = 2$*,* $d = 1$*,* $\Omega = (0, 1)$*, and the following*

[6] We would like to thank the anonymous referee for suggesting this illustrative example.

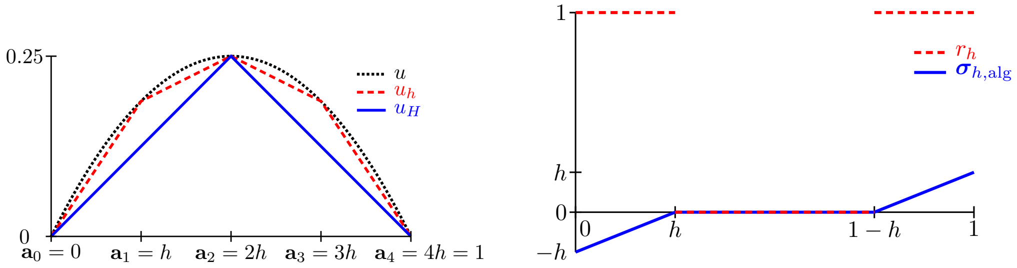


Figure 1: Example 4.6. Setting and exact solution u, approximation u_h on the mesh $\mathcal{T}_h$, and approximation u_H on the twice coarser mesh $\mathcal{T}_H$ (left); r_h from (4.8) and optimal $\boldsymbol{\sigma}_{h,\mathrm{alg}}$ from $\mathbf{RTN}_0$ (right).

situation: $-\Delta u = -u'' = 2$ in Ω and $u = 0$ on $\partial\Omega$, so that the solution of this PDE is $u(x) = x(1-x)$. In the context of (3.11) of Example 3.2, let $V = W_0^{1,2}(\Omega)$, and, for any $u_H \in W_0^{1,2}(\Omega)$, let $\mathcal{R} \in V'$ be defined by

$$\langle \mathcal{R}, v\rangle_{V',V} \coloneqq (2, v) - (\nabla u_H, \nabla v) = \int_0^1 (2v - u_H' v')\,\mathrm{d}x, \qquad v \in V, \tag{4.12}$$

leading to

$$\|\mathcal{R}\|_{V'} = \|\nabla(u - u_H)\|_2 = \left\{ \int_0^1 [(u-u_H)']^2\,\mathrm{d}x \right\}^{\frac{1}{2}}.$$

Let us consider an even integer $N > 0$, define $h \coloneqq 1/N$, and introduce a mesh $\mathcal{T}_h$ of Ω given by the vertices $\boldsymbol{a}_i \coloneqq ih$, $i = 0, \ldots, N$, forming the set $\mathcal{V}_h$ and the elements (intervals) $K_i \coloneqq [\boldsymbol{a}_i, \boldsymbol{a}_{i+1}]$, $i = 0, \ldots, N-1$. We also consider the twice coarser mesh $\mathcal{T}_H$ given similarly by the points $\boldsymbol{a}_{2i} = 2ih$, $i = 0, \ldots, N/2$. Let now u_H be piecewise affine with respect to $\mathcal{T}_H$, $C^0(\overline{\Omega})$-continuous, taking the values of the exact solution u in the vertices $\boldsymbol{a}_i = 2ih$, $i = 0, \ldots, N/2$, see Figure 1, left. This u_H is the finite element solution on the mesh $\mathcal{T}_H$, or, equivalently, the Lagrange interpolate of u on the mesh $\mathcal{T}_H$ (with mesh size $2h$). Consequently,

$$\|\mathcal{R}\|_{V'} = \|\nabla(u - u_H)\|_2 = \mathcal{O}(2h) = \mathcal{O}(h), \tag{4.13}$$

where $g(h) = \mathcal{O}(h)$ when there exist two positive constants c, C independent of h such that $ch \le g(h) \le Ch$ for all $h > 0$. The residual $\mathcal{R}$ generated by the function u_H by (4.12), though, does not satisfy the orthogonality condition (3.20) on $\mathcal{T}_h$. A simple calculation gives

$$\begin{aligned} \langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{V',V} &= (2, \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}} - (\nabla u_H, \nabla \psi_{\boldsymbol{a}})_{\omega_{\boldsymbol{a}}} \\ &= \begin{cases} |\omega_{\boldsymbol{a}}| = 2h & \boldsymbol{a} = \boldsymbol{a}_{2i+1} \in \mathcal{V}_h^{\mathrm{int}} \textit{ odd}, i = 0, \ldots, N/2-1, \\ -|\omega_{\boldsymbol{a}}| = -2h & \boldsymbol{a} = \boldsymbol{a}_{2i} \in \mathcal{V}_h^{\mathrm{int}} \textit{ even}, i = 1, \ldots, N/2-1. \end{cases} \end{aligned} \tag{4.14}$$

Consequently, as $h_\Omega = 1$ and $C_{\mathrm{PF},2,\Omega} = 1/\pi$, $r^{\boldsymbol{a}}$ given by (4.1) take the values $r^{\boldsymbol{a}} = (2h)^{\frac{1}{2}}/\pi$. Thus, since $N_{\mathrm{ov}} = 2$ and $q = 2$,

$$N_{\mathrm{ov}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}} (r^{\boldsymbol{a}})^q \right\}^{\frac{1}{q}} = 2^{\frac{1}{2}} \left\{ \sum_{\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}} (r^{\boldsymbol{a}})^2 \right\}^{\frac{1}{2}} = \frac{2h^{\frac{1}{2}}}{\pi} (N-1)^{\frac{1}{2}} = \mathcal{O}(1). \tag{4.15}$$

Thus by comparison with (4.13), the second term on the right-hand side of (4.2a) critically overestimates $\|\mathcal{R}\|_{V'}$. The same holds for estimate (4.5) with (4.7). Indeed, $C_h^{\overset{\ddagger}{\infty}} = C_{\mathrm{cont,PF}} + \mathcal{O}(h^{-1})$ and consequently (4.2b) and (4.13) give that the right-hand side of (4.5) behaves as $\mathcal{O}(h) + \mathcal{O}(1)$.

Let now u_h be piecewise affine with respect to the mesh $\mathcal{T}_h$, $C^0(\overline{\Omega})$-continuous, taking the values of the exact solution u in the points $\boldsymbol{a}_i = ih$, $i = 0, \ldots, N$. The function u_h is the finite element solution on the mesh $\mathcal{T}_h$, or the Lagrange interpolate of u on the mesh $\mathcal{T}_h$, see

Figure 1, left. (If the residual $\mathcal{R}$ would be defined from u_h and not by (4.12)*, it would satisfy the orthogonality condition* (3.20)*). The triangle inequality gives*

$$\|\mathcal{R}\|_{V'} = \|\nabla(u - u_H)\|_2 \le \|\nabla(u - u_h)\|_2 + \|\nabla(u_h - u_H)\|_2. \tag{4.16}$$

Immediately, $\|\nabla(u - u_h)\|_2 = \mathcal{O}(h)$ and also $\|\nabla(u_h - u_H)\|_2 = \mathcal{O}(h)$, so there is no structural loss in this inequality. Viewing u_H as an approximate solution to u_h, $u_H = \sum_{i=1}^{N-1} u_H(\boldsymbol{a}_i)\psi_{\boldsymbol{a}_i}$, $\mathrm{U}_H \in \mathbb{R}^{N-1}$, $(\mathrm{U}_H)_i = u_H(\boldsymbol{a}_i)$, $i = 1, \ldots, N-1$, where only u_H is supposed to be known explicitly but not u_h, we now consider the most commonly used estimate on the "algebraic" error

$$\|\nabla(u_h - u_H)\|_2^2 = \big(\mathbb{A}_h^{-1}\mathrm{R}_h\big)\cdot\mathrm{R}_h \le \big|\mathbb{A}_h^{-1}\big|_2 |\mathrm{R}_h|_2^2,$$

cf. [45, Section 3.1] and the references therein; here

$$\mathbb{A}_h := \frac{1}{h}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 \end{pmatrix}, \qquad \mathrm{F}_h := \begin{pmatrix} 2h \\ 2h \\ \vdots \\ \vdots \\ 2h \end{pmatrix}$$

are respectively the finite element matrix and the right-hand side vector and

$$\mathrm{R}_h := \mathrm{F}_h - \mathbb{A}_h\mathrm{U}_H = \begin{pmatrix} 2h \\ -2h \\ 2h \\ \vdots \\ 2h \end{pmatrix}$$

is the algebraic residual vector; note that $(\mathrm{R}_h)_i = \langle \mathcal{R}, \psi_{\boldsymbol{a}_i}\rangle_{V',V} = (-1)^{i+1}2h$, $i = 1, \ldots, N-1$, using (4.14)*. Now, cf. [39, Section 7.1] or [45, Section 5.2] and the references therein for similar developments,*

$$|\mathrm{R}_h|_2 = \left\{\sum_{i=1}^{N-1} |(\mathrm{R}_h)_i|^2\right\}^{\frac{1}{2}} = 2h(N-1)^{\frac{1}{2}} = \mathcal{O}(h^{\frac{1}{2}})$$

and

$$\big|\mathbb{A}_h^{-1}\big|_2 = \lambda_{\max}(\mathbb{A}_h^{-1}) = \frac{1}{\lambda_{\min}(\mathbb{A}_h)} = \mathcal{O}(h^{-1}),$$

where the characterization of the smallest eigenvalue $\lambda_{\min}(\mathbb{A}_h) = \mathcal{O}(h)$ of the matrix $\mathbb{A}_h$ in one space dimension is standard, see, e.g., [32, Example 9.15]. Altogether,

$$\|\nabla(u_h - u_H)\|_2 \le \big|\mathbb{A}_h^{-1}\big|_2^{\frac{1}{2}} |\mathrm{R}_h|_2 = \mathcal{O}(1). \tag{4.17}$$

We conclude that the simple estimate of Theorem 4.1 has in this case the same quality as the commonly used l^2-norm estimate of the algebraic residual vector from numerical linear algebra, and that both are greatly imprecise.

Example 4.7 (Optimality of estimate of Theorem 4.3)**.** *We now investigate, for the same setting as in Example 4.6, the quality of the upper bound* (4.10a) *of Theorem 4.3. Following* (4.14)*, the quantities $\frac{1}{|\omega_{\boldsymbol{a}}|}\langle \mathcal{R}, \psi_{\boldsymbol{a}}\rangle_{(V^{\boldsymbol{a}})',V^{\boldsymbol{a}}}$ in* (4.8) *take here the value* 1 *for odd vertices and* -1 *for even vertices. Thus, the elementwise constant function r_h actually vanishes in all the elements except for K_1 and K_N, where it takes the value* 1*, see Figure 1, right. Then, it is easy to check that the best-available $\boldsymbol{\sigma}_{h,\mathrm{alg}}$ from* $\mathbf{RTN}_0$ *such that* $\operatorname{div}\boldsymbol{\sigma}_{h,\mathrm{alg}} = r_h$ *is the function vanishing on all the elements except for K_1 and K_N, depicted in Figure 1, right. This leads to*

$$\|\boldsymbol{\sigma}_{h,\mathrm{alg}}\|_2 = \mathcal{O}(h^{\frac{3}{2}}),$$

The construction of $\boldsymbol{\sigma}_{h,\mathrm{alg}}$ from [39, Section 7.3] then still leads to $\|\boldsymbol{\sigma}_{h,\mathrm{alg}}\|_2 = \mathcal{O}(h^{\frac{3}{2}})$, whereas that from [44, Definition 6.3] yields $\|\boldsymbol{\sigma}_{h,\mathrm{alg}}\|_2 = \mathcal{O}(h)$. Consequently, in both practical constructions of Example 4.4, the second term on the right-hand side of (4.10a) *does not spoil the quality of the estimate, in contrast to* (4.2a) *with* (4.15) *and* (4.16) *with* (4.17)*.*

Having identified the additional terms in inequalities (4.2a) and (4.10a), one typically controls adaptively their size of with respect to the principal contribution $\left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}}$ (and similarly for $\max_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}$ if $p=1$), see, e.g., [39, equation (6.1)] or [33, equation (3.10)]. The following corollary shows that *localization of* $\|\mathcal{R}\|_{V'}$ can be *restored* in this way. It, however, follows from Examples 4.6 and 4.7 that it may be excessively costly to satisfy the balance condition (4.18) in the case of Theorem 4.1, in contrast to the case of Theorem 4.3.

Corollary 4.8 (Localization of dual norms of functionals with controlled loss of orthogonality)**.** *Let $\mathcal{R}\in V'$ be arbitrary, and consider either the context of Theorem 4.1 with*

$$r_{\mathrm{res}} := \left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}}(r^{\boldsymbol{a}})^q\right\}^{\frac{1}{q}} \quad \textit{if } 1<p\leq\infty, \qquad r_{\mathrm{res}} := \max_{\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{int}}} r^{\boldsymbol{a}} \quad \textit{if } p=1,$$

or the context of Theorem 4.3 with

$$r_{\mathrm{res}} := \frac{1}{N_{\mathrm{ov}}}\|\boldsymbol{\sigma}_{h,\mathrm{alg}}\|_q.$$

Assume moreover that

$$r_{\mathrm{res}} \leq \gamma_{\mathrm{res}} C_{\mathrm{cont,PF}}\left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}}, \qquad 1<p\leq\infty, \tag{4.18a}$$

$$r_{\mathrm{res}} \leq \gamma_{\mathrm{res}} C_{\mathrm{cont,PF}}\max_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}, \qquad p=1 \tag{4.18b}$$

for some parameter $\gamma_{\mathrm{res}}\geq 0$. Then, when $1<p\leq\infty$,

$$\|\mathcal{R}\|_{V'} \leq (1+\gamma_{\mathrm{res}})N_{\mathrm{ov}}C_{\mathrm{cont,PF}}\left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}},$$

$$\left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}} \leq \|\mathcal{R}\|_{V'},$$

and, when $p=1$,

$$\|\mathcal{R}\|_{V'} \leq (1+\gamma_{\mathrm{res}})N_{\mathrm{ov}}C_{\mathrm{cont,PF}}\max_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'},$$

$$\max_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \leq \|\mathcal{R}\|_{V'}.$$

4.2 Localization in vectorial setting

We now finally present a vectorial variant of Theorem 3.7, with typical applications in Stokes-type fluid flows, cf. [14]. We only make a concise presentation, as the extension from the scalar case is rather straightforward.

Let $\nabla\boldsymbol{v}$ for $\boldsymbol{v}\in[W^{1,p}(\omega)]^d$ be the matrix with lines given by $\nabla\boldsymbol{v}_i$, $1\leq i\leq d$; in accordance with the notation of Section 2.1, $\|\nabla\boldsymbol{v}\|_{p,\omega} := (\int_\omega(\sum_{i=1}^d\sum_{j=1}^d|\partial_{\boldsymbol{x}_j}\boldsymbol{v}_i(\boldsymbol{x})|^2)^{\frac{p}{2}}\,\mathrm{d}\boldsymbol{x})^{\frac{1}{p}}$. For vectors $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d$, $\boldsymbol{u}\otimes\boldsymbol{v}$ defines a tensor $\mathbb{T}\in\mathbb{R}^{d\times d}$ such that $\mathbb{T}_{i,j} := \boldsymbol{u}_i\boldsymbol{v}_j$. Then the vectorial variant of Lemma 2.4 is:

Lemma 4.9 (Cut-off estimate in vectorial setting)**.** *There exists a constant $C_{\mathrm{cont,PF},d}>0$, only depending on the space dimension d and on the constant $C_{\mathrm{cont,PF}}$ from* (2.6)*, such that for all $\boldsymbol{a}\in\mathcal{V}_h$, there holds*

$$\|\nabla(\psi_{\boldsymbol{a}}\boldsymbol{v})\|_{p,\omega_{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF},d}\|\nabla\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}} \qquad \forall\boldsymbol{v}\in[W^{1,p}_*(\omega_{\boldsymbol{a}})]^d.$$

Proof. Assume first $1 \le p < \infty$. Using the scalar Poincaré–Friedrichs inequality (2.5) and the norm equivalence (2.1),

$$\begin{aligned}\|\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}} &= \left(\int_{\omega_{\boldsymbol{a}}}\left(\sum_{i=1}^{d}|\boldsymbol{v}_i(\boldsymbol{x})|^2\right)^{\frac{p}{2}}\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}} \le \underline{C}_{p,d}\left(\sum_{i=1}^{d}\int_{\omega_{\boldsymbol{a}}}|\boldsymbol{v}_i(\boldsymbol{x})|^p\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}}\\ &\le \underline{C}_{p,d}C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}}h_{\omega_{\boldsymbol{a}}}\left(\sum_{i=1}^{d}\|\nabla\boldsymbol{v}_i\|^p_{p,\omega_{\boldsymbol{a}}}\right)^{\frac{1}{p}} \le \underline{C}_{p,d}\overline{C}_{p,d}C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}}h_{\omega_{\boldsymbol{a}}}\|\nabla\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\boldsymbol{v}\in[W^{1,p}_*(\omega_{\boldsymbol{a}})]^d,\end{aligned}$$

where

$$\underline{C}_{p,d} := \begin{cases} 1 & \text{if } p\le 2,\\ d^{\frac{1}{2}-\frac{1}{p}} & \text{if } p \ge 2\end{cases}$$

and

$$\overline{C}_{p,d} := \begin{cases} d^{\frac{1}{p}-\frac{1}{2}} & \text{if } p\le 2,\\ 1 & \text{if } p \ge 2.\end{cases}$$

Denote $C_{p,d} := \underline{C}_{p,d}\overline{C}_{p,d} = d^{\left|\frac{1}{2}-\frac{1}{p}\right|}$ and notice that $1 \le C_{p,d} \le \sqrt{d}$. Then, we readily arrive at

$$\begin{aligned}\|\nabla(\psi_{\boldsymbol{a}}\boldsymbol{v})\|_{p,\omega_{\boldsymbol{a}}} &= \|\boldsymbol{v}\otimes\nabla\psi_{\boldsymbol{a}} + \psi_{\boldsymbol{a}}\nabla\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}}\\ &\le \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}\|\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}} + \|\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}\|\nabla\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}}\\ &\le (1 + C_{p,d}C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}}h_{\omega_{\boldsymbol{a}}}\|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}})\|\nabla\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}},\end{aligned}$$

and the assertion follows with $C_{\mathrm{cont,PF},d} := \max_{\boldsymbol{a}\in\mathcal{V}_h}\{1 + C_{p,d}C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}}h_{\omega_{\boldsymbol{a}}}\|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}\}$. Case $p = \infty$ is an obvious modification. □

Denote

$$\begin{aligned}\boldsymbol{V} &:= [W^{1,p}_0(\Omega)]^d,\\ \mathcal{R} &\in \boldsymbol{V}',\\ \|\mathcal{R}\|_{\boldsymbol{V}'} &:= \sup_{\boldsymbol{v}\in\boldsymbol{V};\,\|\nabla\boldsymbol{v}\|_p=1}\langle\mathcal{R},\boldsymbol{v}\rangle_{\boldsymbol{V}',\boldsymbol{V}}.\end{aligned}$$

For a vertex $\boldsymbol{a}\in\mathcal{V}_h$, let the local setting be

$$\begin{aligned}\boldsymbol{V}^{\boldsymbol{a}} &:= [W^{1,p}_0(\omega_{\boldsymbol{a}})]^d,\\ \langle\mathcal{R},\boldsymbol{v}\rangle_{(\boldsymbol{V}^{\boldsymbol{a}})',\boldsymbol{V}^{\boldsymbol{a}}} &:= \langle\mathcal{R},\boldsymbol{v}\rangle_{\boldsymbol{V}',\boldsymbol{V}} \qquad \boldsymbol{v}\in\boldsymbol{V}^{\boldsymbol{a}},\\ \|\mathcal{R}\|_{(\boldsymbol{V}^{\boldsymbol{a}})'} &:= \sup_{\boldsymbol{v}\in\boldsymbol{V}^{\boldsymbol{a}};\,\|\nabla\boldsymbol{v}\|_{p,\omega_{\boldsymbol{a}}}=1}\langle\mathcal{R},\boldsymbol{v}\rangle_{(\boldsymbol{V}^{\boldsymbol{a}})',\boldsymbol{V}^{\boldsymbol{a}}}.\end{aligned}$$

Define $\boldsymbol{\psi}_{\boldsymbol{a},m}$, $1\le m\le d$, as the vectorial variant of the partition of unity functions $\psi_{\boldsymbol{a}}$ such that $(\boldsymbol{\psi}_{\boldsymbol{a},m})_m = \psi_{\boldsymbol{a}}$ and $(\boldsymbol{\psi}_{\boldsymbol{a},m})_n = 0$ for $1\le n\le d$, $n\ne m$. The following is a generalization of Theorem 3.7 to vectorial setting:

Theorem 4.10 (Localization of dual norms of functionals in vectorial case)**.** *Let $\mathcal{R}\in\boldsymbol{V}'$ be arbitrary and let*

$$\langle\mathcal{R},\boldsymbol{\psi}_{\boldsymbol{a},m}\rangle_{\boldsymbol{V}',\boldsymbol{V}} = 0 \qquad \forall 1\le m\le d,\,\forall\boldsymbol{a}\in\mathcal{V}^{\mathrm{int}}_h.$$

Then, when $1 < p \le \infty$,

$$\|\mathcal{R}\|_{\boldsymbol{V}'} \le N_{\mathrm{ov}}C_{\mathrm{cont,PF},d}\left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|^q_{(\boldsymbol{V}^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}},$$

$$\left\{\frac{1}{N_{\mathrm{ov}}}\sum_{\boldsymbol{a}\in\mathcal{V}_h}\|\mathcal{R}\|^q_{(\boldsymbol{V}^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}} \le \|\mathcal{R}\|_{\boldsymbol{V}'},$$

and, when $p = 1$,

$$\begin{aligned}\|\mathcal{R}\|_{\boldsymbol{V}'} &\le N_{\mathrm{ov}} C_{\mathrm{cont,PF},d} \max_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|_{(\boldsymbol{V}^{\boldsymbol{a}})'},\\ \max_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|_{(\boldsymbol{V}^{\boldsymbol{a}})'} &\le \|\mathcal{R}\|_{\boldsymbol{V}'}.\end{aligned}$$

The orthogonality condition is actually again only needed in the first inequalities.

Proof. Along the lines of proof of Theorem 3.7, using Lemma 4.9 instead of Lemma 2.4. □

Extension of Remark 3.10, Theorems 4.1 and 4.3, and of Corollary 4.8 to vectorial case is straightforward.

5 Numerical illustration

We now numerically demonstrate the validity of Theorem 3.7 in the setting $1 < p < \infty$. The experiments were implemented using `dolfin-tape` [11] package built on top of the FEniCS Project [2]. The complete supporting code for reproducing the experiments can be obtained at [12].

Let $V_h := \mathbb{P}_1(\mathcal{T}_h) \cap W^{1,p}(\Omega)$ be the space of continuous, piecewise first-order polynomials with respect to a matching triangular mesh $\mathcal{T}_h$ of the domain $\Omega \subset \mathbb{R}^2$, see Remark 2.1. Let $V_h^0 := V_h \cap W_0^{1,p}(\Omega)$ be its zero-trace subspace and let u_h be a finite element approximation to the p-Laplace problem (3.11) of Example 3.2, i.e.,

$$u_h - u_h^{\mathrm{D}} \in V_h^0, \tag{5.1a}$$

$$(|\nabla u_h|^{p-2}\nabla u_h, \nabla v_h) = (f_h, v_h) \qquad \forall v_h \in V_h^0, \tag{5.1b}$$

where $u_h^{\mathrm{D}} \in V_h$ is a $\mathbb{P}_1$-nodal interpolant of $u^{\mathrm{D}} \in W^{1,p}(\Omega) \cap C^0(\overline{\Omega})$ (approximation error of u^{D} by u_h^{D} is neglected) and $(f_h, \cdot)$ approximates $(f, \cdot)$ by a six-node quadrature rule with fourth-order precision from [47, p. 184, Table 4.1]. We consider $\mathcal{R} \in V'$, the residual of u_h with respect to equation (3.11b) (with $\boldsymbol{\sigma}(\nabla u) = |\nabla u|^{p-2}\nabla u$) given by (3.12). Taking $v_h = \psi_{\boldsymbol{a}}$ in (5.1b) immediately gives the orthogonality property (3.20) for all interior vertices $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$. Computationally, regularization and linearization of the degenerate p-Laplace operator is employed to approximately solve (5.1). The arising errors are secured to be small by error-distinguishing a posteriori estimation techniques of [33], thus ensuring sufficiently approximate fulfillment of the Galerkin orthogonality (3.20).

The evaluation of the norms $\|\mathcal{R}\|_{V'}$ and $\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}$ in (3.21)–(3.22) is equivalent to solving respectively for the *global lifting* $\mathscr{r}$ on Ω defined by (3.17) and for the *local liftings* $\mathscr{r}^{\boldsymbol{a}}$ on every patch $\omega_{\boldsymbol{a}}$ defined by (3.28). Again, only approximations $\mathscr{r}_h \in V$ and $\mathscr{r}_h^{\boldsymbol{a}} \in V^{\boldsymbol{a}}$ are available, where the evaluation error $\mathcal{E}_h \in V'$ is given by

$$\langle \mathcal{E}_h, v\rangle_{V',V} := (|\nabla \mathscr{r}_h|^{p-2}\nabla \mathscr{r}_h, \nabla v) - \langle \mathcal{R}, v\rangle_{V',V} \qquad v \in V.$$

Since, simultaneously,

$$\begin{aligned}\|\mathcal{R}\|_{V'} &\le \|\mathcal{E}_h\|_{V'} + \|\nabla \mathscr{r}_h\|_p^{p-1},\\ \|\nabla \mathscr{r}_h\|_p^{p-1} &\le \|\mathcal{R}\|_{V'} + \|\mathcal{E}_h\|_{V'},\end{aligned}$$

we obtain

$$\frac{\big|\|\nabla \mathscr{r}_h\|_p^{p-1} - \|\mathcal{R}\|_{V'}\big|}{\|\nabla \mathscr{r}_h\|_p^{p-1}} \le \frac{\|\mathcal{E}_h\|_{V'}}{\|\nabla \mathscr{r}_h\|_p^{p-1}}.$$

Consequently, using a posteriori techniques from [33], the approximation

$$\|\mathcal{R}\|_{V'} \approx \|\nabla \mathscr{r}_h\|_p^{p-1}$$

is guaranteed to hold with a given relative accuracy that we set to 10^{-2}. Similarly, we secure the relative accuracy of the approximation

$$\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \approx \|\nabla \mathscr{r}_h^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}^{p-1}$$

to 10^{-2}. For clarity of notation, we drop the subscript h in what follows.

In order to plot local distributions, we find it natural to define two non-negative functions from $\mathbb{P}_1(\mathcal{T}_h)$

$$\epsilon_{\text{glob}}^q := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla \mathscr{r}\|_{p,\omega_{\boldsymbol{a}}}^p \frac{\psi_{\boldsymbol{a}}}{|\omega_{\boldsymbol{a}}|}, \tag{5.2a}$$

$$\epsilon_{\text{loc}}^q := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla \mathscr{r}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}^p \frac{\psi_{\boldsymbol{a}}}{|\omega_{\boldsymbol{a}}|}. \tag{5.2b}$$

The employed normalization gives on simplicial meshes $|\omega_{\boldsymbol{a}}|^{-1}(\psi_{\boldsymbol{a}}, 1)_{\omega_{\boldsymbol{a}}} = N_{\text{ov}}^{-1}$ (with $N_{\text{ov}} = d+1$) and together with (2.7a) ensures that

$$\begin{aligned} \|\epsilon_{\text{glob}}\|_q^q &= \|\nabla \mathscr{r}\|_p^p \overset{(3.18)}{=} \|\mathcal{R}\|_{V'}^q, \\ \|\epsilon_{\text{loc}}\|_q^q &= \frac{1}{N_{\text{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla \mathscr{r}^{\boldsymbol{a}}\|_{p,\omega_{\boldsymbol{a}}}^p \overset{(3.27)}{=} \frac{1}{N_{\text{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}^q. \end{aligned}$$

Consequently, Theorem 3.7 can be rephrased as

$$\begin{aligned} \|\epsilon_{\text{glob}}\|_q &\le N_{\text{ov}} C_{\text{cont,PF}} \|\epsilon_{\text{loc}}\|_q, \\ \|\epsilon_{\text{loc}}\|_q &\le \|\epsilon_{\text{glob}}\|_q. \end{aligned}$$

Moreover, the second inequality above can be split into local contributions using Remark 3.10, so that

$$\epsilon_{\text{loc}} \le \epsilon_{\text{glob}}. \tag{5.3}$$

Let us also introduce the effectivity index of an inequality (`ineq`)

$$\text{Eff}_{(\texttt{ineq})} := \frac{\text{rhs of } (\texttt{ineq})}{\text{lhs of } (\texttt{ineq})} \ge 1.$$

For testing, we choose

- *Chaillou–Suri* [25, 33], $\Omega = (0,1)^2$, $p \in \{1.5, 10\}$, $u^{\text{D}}(\mathbf{x}) = q^{-1}(0.5^q - |\mathbf{x} - (0.5, 0.5)|^q)$, $f = -\Delta_p u^{\text{D}} = 2$,
- *Carstensen–Klose* [24, Example 3], $\Omega = (-1,1)^2 \setminus [0,1] \times [-1,0]$, $p = 4$, $u^{\text{D}}(r,\theta) = r^{\frac{7}{8}} \sin(\frac{7}{8}\theta)$, $f = -\Delta_p u^{\text{D}}$.

As we have the exact solution $u = u^{\text{D}}$ in our hands, we can also check the distribution of the flux error (3.14) and of the $W_0^{1,p}(\Omega)$-norm error (3.15). Therefore, as above, we define non-negative functions from $\mathbb{P}_1(\mathcal{T}_h)$

$$\epsilon_{\text{flux}}^q := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_{q,\omega_{\boldsymbol{a}}}^q \frac{\psi_{\boldsymbol{a}}}{|\omega_{\boldsymbol{a}}|}, \tag{5.4}$$

$$\epsilon_{\text{en}}^p := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \|\nabla(u - u_h)\|_{p,\omega_{\boldsymbol{a}}}^p \frac{\psi_{\boldsymbol{a}}}{|\omega_{\boldsymbol{a}}|} \tag{5.5}$$

having properties

$$\begin{aligned} \|\epsilon_{\text{flux}}\|_q &= \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q, \\ \|\epsilon_{\text{en}}\|_p &= \|\nabla(u - u_h)\|_p. \end{aligned}$$

Estimates (3.13) translate to

$$\|\epsilon_{\text{glob}}\|_q \le \|\epsilon_{\text{flux}}\|_q, \tag{5.6a}$$

$$\epsilon_{\text{loc}} \le \epsilon_{\text{flux}}. \tag{5.6b}$$

The results of numerical experiments are shown in Table 1 and Figures 2–5. Effectivity indices in Table 1 show that the reverse bound (3.21b) is quite tight but the forward bound (3.21a)

Table 1: Computed quantities of localization inequalities (3.21), (5.7), and of estimate (3.13a) for the chosen model problems. Recall that $\|\epsilon_{\mathrm{glob}}\|_q = \|\mathcal{R}\|_{V'}$, $\|\epsilon_{\mathrm{loc}}\|_q = \left\{\sum_{\boldsymbol{a}\in\mathcal{V}_h} \frac{1}{N_{\mathrm{ov}}}\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'}\right\}^{\frac{1}{q}}$, and $\|\epsilon_{\mathrm{flux}}\|_q = \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q$.

Case	#cells	$C_{\mathrm{cont,PF}}$	$\|\epsilon_{\mathrm{glob}}\|_q$	$\|\epsilon_{\mathrm{loc}}\|_q$	$\|\epsilon_{\mathrm{flux}}\|_q$	$\mathrm{Eff}_{(3.21\mathrm{a})}$	$\mathrm{Eff}_{(5.7)}$	$\mathrm{Eff}_{(3.21\mathrm{b})}$	$\mathrm{Eff}_{(3.13\mathrm{a})}$
Chaillou–Suri $p = 1.5$, $N_{\mathrm{ov}} = 3$	100	5.670	0.0502	0.0431	0.0546	14.6	13.8	1.17	1.09
	400	5.670	0.0259	0.0220	0.0274	14.4	14.1	1.18	1.06
	900	5.670	0.0174	0.0147	0.0183	14.4	14.2	1.18	1.05
	1600	5.670	0.0131	0.0111	0.0137	14.4	14.2	1.18	1.04
Chaillou–Suri $p = 10.0$, $N_{\mathrm{ov}} = 3$	100	7.645	0.0604	0.0484	0.1043	18.4	16.6	1.25	1.73
	400	7.645	0.0312	0.0255	0.0501	18.8	17.8	1.22	1.61
	900	7.645	0.0214	0.0175	0.0343	18.8	18.1	1.22	1.60
	1600	7.645	0.0161	0.0132	0.0255	18.8	18.4	1.22	1.58
Carstensen–Klose $p = 4.0$, $N_{\mathrm{ov}} = 3$	40	9.706	0.1611	0.1236	0.1889	22.3	16.3	1.30	1.17
	189	13.844	0.0930	0.0753	0.1029	33.6	19.0	1.23	1.11
	428	12.981	0.0635	0.0518	0.0701	31.8	19.4	1.23	1.10
	739	12.801	0.0471	0.0383	0.0527	31.2	19.9	1.23	1.12

suffers by a larger, though still reasonable and predictable, overestimation. This overestimation decreases a little when improving (3.21a) to

$$\|\mathcal{R}\|_{V'} \leq N_{\mathrm{ov}} \left\{ \frac{1}{N_{\mathrm{ov}}} \sum_{\boldsymbol{a}\in\mathcal{V}_h} \big(C_{\mathrm{cont,PF},\omega_{\boldsymbol{a}}} \|\mathcal{R}\|_{(V^{\boldsymbol{a}})'} \big)^q \right\}^{\frac{1}{q}}, \tag{5.7}$$

where $C_{\mathrm{cont,PF},\omega_{\boldsymbol{a}}} := 1 + C_{\mathrm{PF},p,\omega_{\boldsymbol{a}}} h_{\omega_{\boldsymbol{a}}} \|\nabla\psi_{\boldsymbol{a}}\|_{\infty,\omega_{\boldsymbol{a}}}$ is the continuity constant of each patch; this improvement is much more significant for the case with singularity (*Carstensen–Klose*), see Table 1; we conjecture that the improvement would lose its significance if the residuals $\mathcal{R}$ were obtained on a sequence of adaptively refined meshes.

Figures 2, 3, 4, and 5 nicely demonstrate the local inequalities (3.29) and (3.13b) as expressed by (5.3) and (5.6b), respectively. The figures also show that there is no hope of locally comparing the $W_0^{1,p}(\Omega)$-norm error $\|\nabla(u-u_h)\|_p^p$ (expressed here by ϵ_{en}^p of (5.5)) and the lifted residual error $\|\nabla \mathscr{r}\|_p^p$ (expressed here by $\epsilon_{\mathrm{glob}}^q$ of (5.2a)). The colorbars systematically present the minimal and maximal values, taken at the vertices $\boldsymbol{a} \in \mathcal{V}_h$. In the plots, there is one color per triangle, corresponding to the mean value over its vertices.

In conclusion, the main result, Theorem 3.7, as well as Remark 3.10, are well supported by the performed numerical experiments.

Acknowledgement

The authors would like to thank the anonymous referees for stimulating suggestions that have helped to enrich the original manuscript and Iain Smears (University College London) for interesting discussions on the subject.

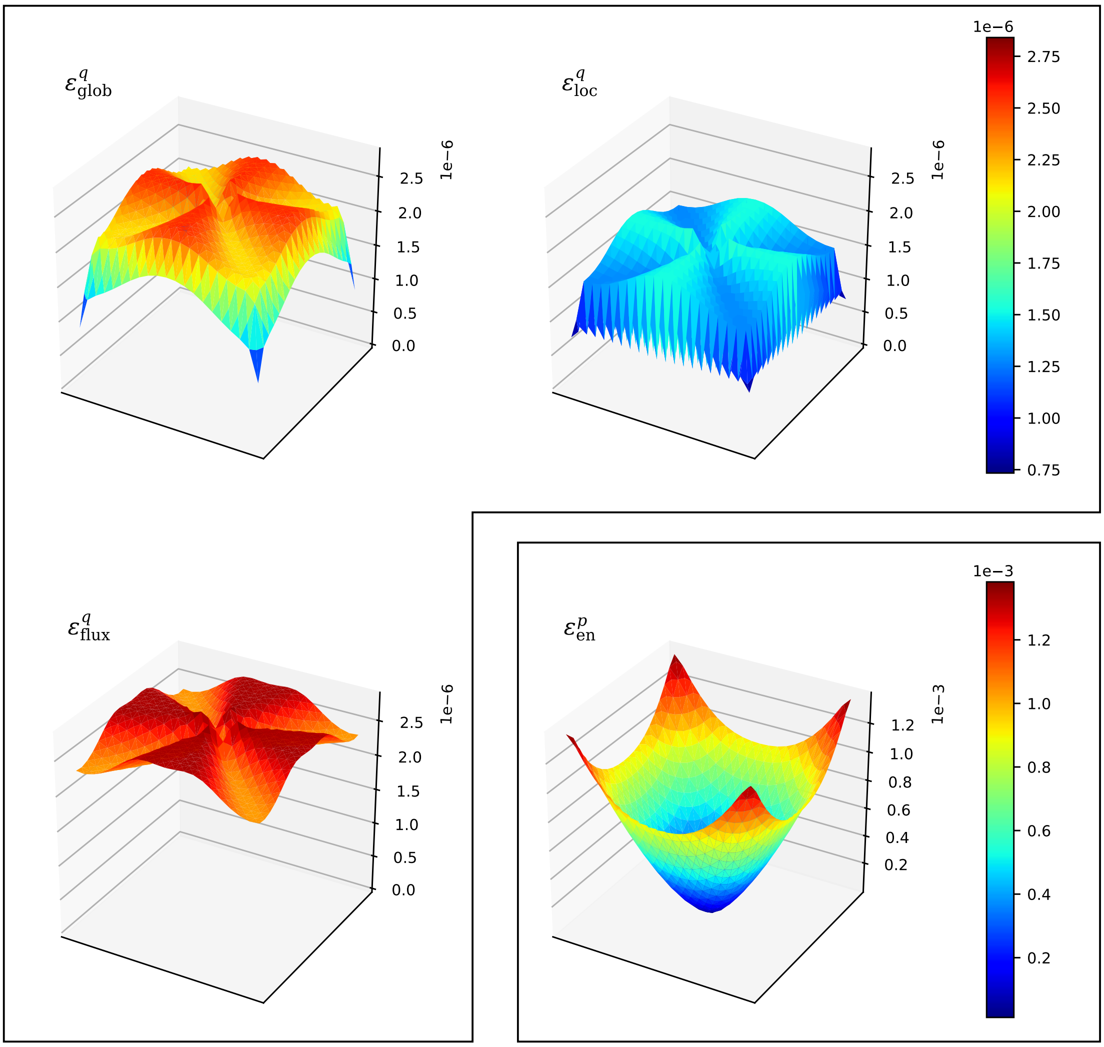


Figure 2: Distribution of $\|\epsilon_{\text{glob}}\|_q^q = \|\mathcal{R}\|_{V'}^q$ (top left), $\|\epsilon_{\text{loc}}\|_q^q = \sum_{\boldsymbol{a}\in\mathcal{V}_h} \frac{1}{N}\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}^q$ (top right), $\|\epsilon_{\text{flux}}\|_q^q = \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q^q$ (bottom left), and $\|\epsilon_{\text{en}}\|_p^p = \|\nabla(u - u_h)\|_p^p$ (bottom right) for the case *Chaillou–Suri*, $p = 1.5$, #cells=1600

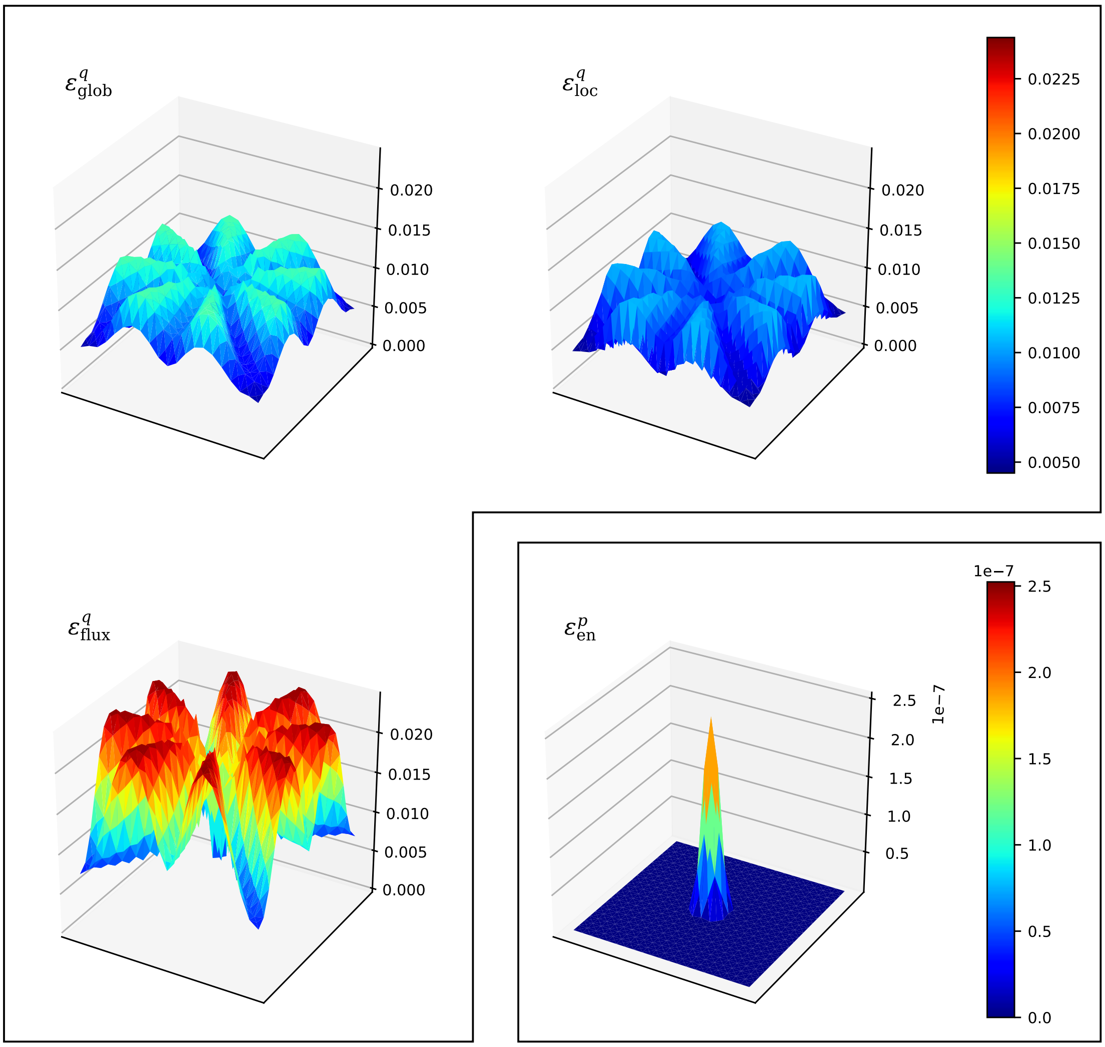


Figure 3: Distribution of $\|\epsilon_{\mathrm{glob}}\|_q^q = \|\mathcal{R}\|_{V'}^q$ (top left), $\|\epsilon_{\mathrm{loc}}\|_q^q = \sum_{\boldsymbol{a}\in\mathcal{V}_h} \frac{1}{N}\|\mathcal{R}\|_{(V^{\boldsymbol{a}})'}^q$ (top right), $\|\epsilon_{\mathrm{flux}}\|_q^q = \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q^q$ (bottom left), and $\|\epsilon_{\mathrm{en}}\|_p^p = \|\nabla(u - u_h)\|_p^p$ (bottom right) for the case *Chaillou–Suri*, $p = 10$, #cells=1600

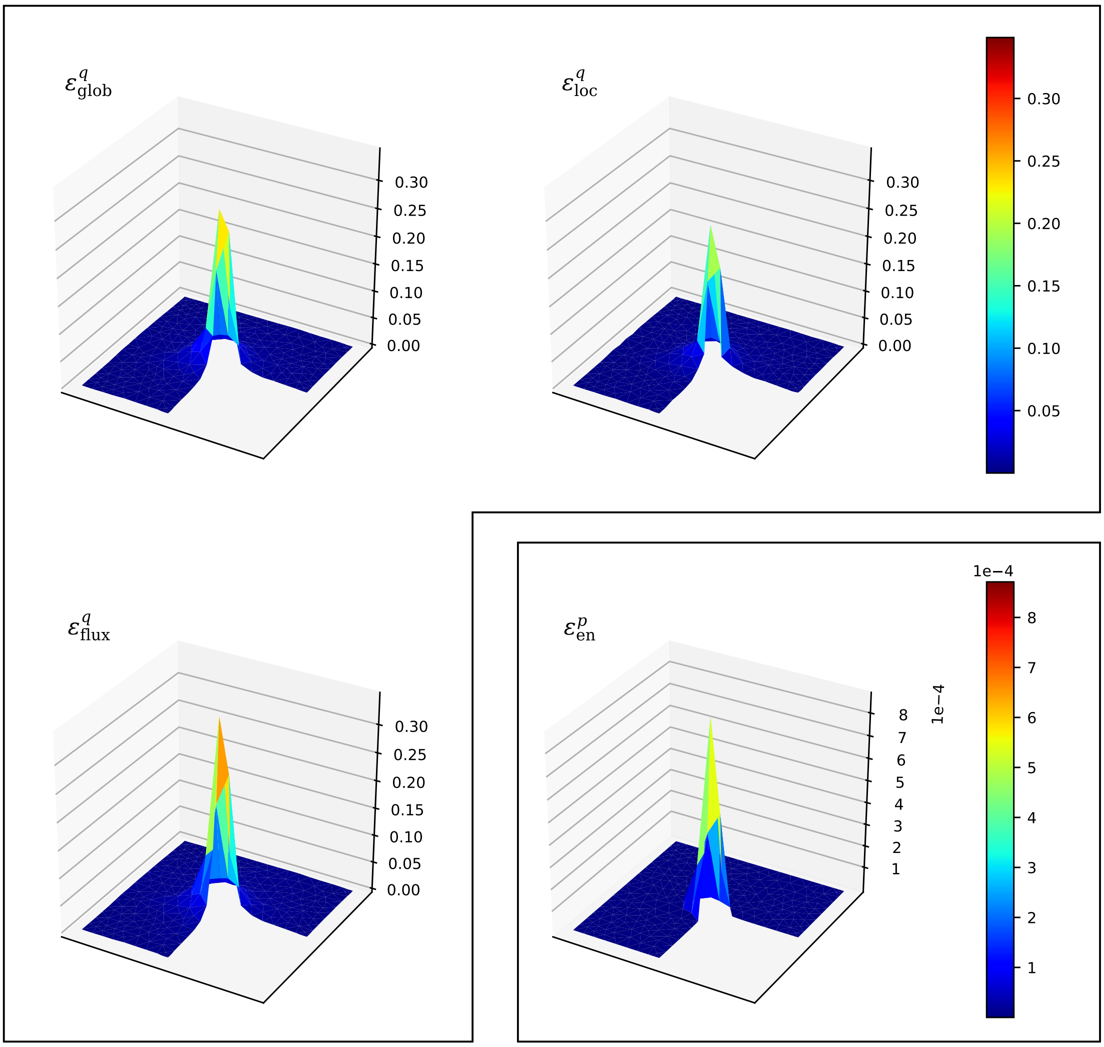


Figure 4: Distribution of $\|\epsilon_{\rm glob}\|_q^q = \|\mathcal{R}\|_{V'}^q$ (top left), $\|\epsilon_{\rm loc}\|_q^q = \sum_{\boldsymbol{a}\in\mathcal{V}_h} \frac{1}{N}\|\mathcal{R}\|^q_{(V^{\boldsymbol{a}})'}$ (top right), $\|\epsilon_{\rm flux}\|_q^q = \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|_q^q$ (bottom left), and $\|\epsilon_{\rm en}\|_p^p = \|\nabla(u - u_h)\|_p^p$ (bottom right) for the case *Carstensen–Klose*, $p = 4$, #cells=428

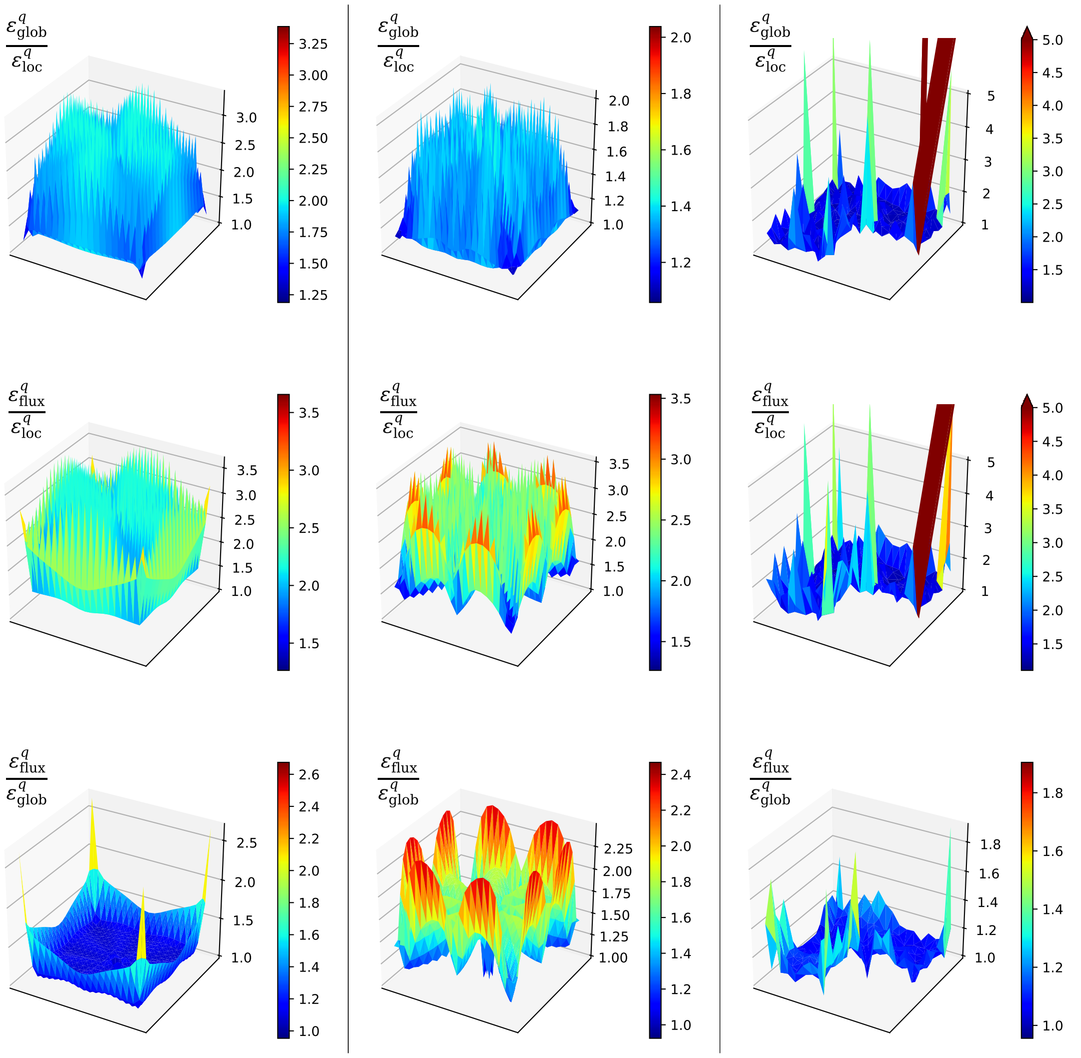


Figure 5: Local ratios of error distributions ϵ^q_{glob}, ϵ^q_{loc}, and ϵ^q_{flux} as functions $\sum_{\boldsymbol{a}\in\mathcal{V}_h} \alpha_{\boldsymbol{a}} \frac{\psi_{\boldsymbol{a}}}{|\omega_{\boldsymbol{a}}|}$ from $\mathbb{P}_1(\mathcal{T}_h)$ with respectively $\alpha_{\boldsymbol{a}} = \|\nabla \mathscr{r}\|^p_{p,\omega_{\boldsymbol{a}}} / \|\nabla \mathscr{r}^{\boldsymbol{a}}\|^p_{p,\omega_{\boldsymbol{a}}}$, $\alpha_{\boldsymbol{a}} = \|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|^q_{q,\omega_{\boldsymbol{a}}} / \|\nabla \mathscr{r}^{\boldsymbol{a}}\|^p_{p,\omega_{\boldsymbol{a}}}$, and $\|\boldsymbol{\sigma}(\nabla u) - \boldsymbol{\sigma}(\nabla u_h)\|^q_{q,\omega_{\boldsymbol{a}}} / \|\nabla \mathscr{r}\|^p_{p,\omega_{\boldsymbol{a}}}$ for cases *Chaillou–Suri*, $p = 1.5$, #cells=1600 (left), *Chaillou–Suri*, $p = 10$, #cells=1600 (middle), *Carstensen–Klose*, $p = 4$, #cells=428 (right). The top and middle row express effectivity of inequalities (5.3) and (5.6b) respectively, hence the quantities are bounded from below by one; the bottom quantity is not known to be bounded from below by one and did not turn out to be bounded in the experiments.

References

[1] G. Acosta and R. G. Durán. “An optimal Poincaré inequality in L^1 for convex domains”. In: *Proc. Amer. Math. Soc.* 132.1 (2004), pp. 195–202. DOI: 10.1090/S0002-9939-03-07004-7.

[2] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. Rognes, and G. Wells. “The FEniCS Project Version 1.5”. In: *Archive of Numerical Software* 3.100 (2015). DOI: 10.11588/ans.2015.100.20553.

[3] M. Aurada, M. Feischl, J. Kemetmüller, M. Page, and D. Praetorius. “Each $H^{1/2}$-stable projection yields convergence and quasi-optimality of adaptive FEM with inhomogeneous Dirichlet data in $\mathbb{R}^d$”. In: *ESAIM Math. Model. Numer. Anal.* 47.4 (2013), pp. 1207–1235. DOI: 10.1051/m2an/2013069.

[4] I. Babuška and J. M. Melenk. “The partition of unity method”. In: *Internat. J. Numer. Methods Engrg.* 40.4 (1997), pp. 727–758. DOI: 10.1002/(SICI)1097-0207(19970228)40:4<727::AID-NME86>3.3.CO;2-E.

[5] I. Babuška and A. Miller. “A feedback finite element method with a posteriori error estimation. I. The finite element method and some basic properties of the a posteriori error estimator”. In: *Comput. Methods Appl. Mech. Engrg.* 61.1 (1987), pp. 1–40. DOI: 10.1016/0045-7825(87)90114-9.

[6] J. W. Barrett and W. B. Liu. “Finite element approximation of the p-Laplacian”. In: *Math. Comp.* 61.204 (1993), pp. 523–537.

[7] J. W. Barrett and W. B. Liu. “Quasi-norm error bounds for the finite element approximation of a non-Newtonian flow”. In: *Numer. Math.* 68.4 (1994), pp. 437–456. DOI: 10.1007/s002110050071.

[8] Y. Bazilevs, L. Beirão da Veiga, J. A. Cottrell, T. J. R. Hughes, and G. Sangalli. “Isogeometric analysis: approximation, stability and error estimates for h-refined meshes”. In: *Math. Models Methods Appl. Sci.* 16.7 (2006), pp. 1031–1090. DOI: 10.1142/S0218202506001455.

[9] L. Beck, M. Bulíček, J. Málek, and E. Süli. “On the existence of integrable solutions to nonlinear elliptic systems and variational problems with linear growth”. In: *Arch. Ration. Mech. Anal.* 225.2 (2017), pp. 717–769. DOI: 10.1007/s00205-017-1113-4.

[10] L. Belenki, L. Diening, and C. Kreuzer. “Optimality of an adaptive finite element method for the p-Laplacian equation”. In: *IMA J. Numer. Anal.* 32.2 (2012), pp. 484–510. DOI: 10.1093/imanum/drr016.

[11] J. Blechta. *dolfin-tape, DOLFIN tools for a posteriori error estimation, version "paper-norms-nonlin-code-v1.0-rc3"*. June 2016. DOI: 10.5281/zenodo.55443.

[12] J. Blechta. *Supporting code for “Localization of the $W^{-1,q}$ norm for local a posteriori efficiency”*. June 2018. DOI: 10.5281/zenodo.1302993.

[13] J. Blechta, J. Málek, and K. Rajagopal. “On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion”. arXiv preprint arXiv:1902.04853v1, submitted for publication. 2019. URL: https://arxiv.org/abs/1902.04853v1.

[14] J. Blechta, J. Málek, and M. Vohralík. “Generalized Stokes flows of implicitly constituted fluids: a posteriori error control and full adaptivity”. In preparation. 2019.

[15] D. Braess, V. Pillwein, and J. Schöberl. “Equilibrated residual error estimates are p-robust”. In: *Comput. Methods Appl. Mech. Engrg.* 198.13-14 (2009), pp. 1189–1197. DOI: 10.1016/j.cma.2008.12.010.

[16] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Third. Vol. 15. Texts in Applied Mathematics. Springer, New York, 2008, pp. xviii+397. DOI: 10.1007/978-0-387-75934-0.

[17] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Vol. 15. Springer Series in Computational Mathematics. New York: Springer-Verlag, 1991, pp. x+350. DOI: 10.1007/978-1-4612-3172-1.

[18] A. Buffa and C. Giannelli. “Adaptive isogeometric methods with hierarchical splines: error estimator and convergence”. In: *Math. Models Methods Appl. Sci.* 26.1 (2016), pp. 1–25. DOI: 10.1142/S0218202516500019.

[19] M. Bulíček, P. Gwiazda, J. Málek, K. R. Rajagopal, and A. Świerczewska-Gwiazda. “On flows of fluids described by an implicit constitutive equation characterized by a maximal monotone graph”. In: *Mathematical Aspects of Fluid Mechanics*. Ed. by J. C. Robinson, J. L. Rodrigo, and W. Sadwoski. Vol. 402. LMS Lecture Notes Series. Cambridge, Great Britain: Cambridge University Press, 2012, pp. 23–51.

[20] M. Bulíček, P. Gwiazda, J. Málek, and A. Świerczewska-Gwiazda. “On unsteady flows of implicitly constituted incompressible fluids”. In: *SIAM J. Math. Anal.* 44.4 (2012), pp. 2756–2801. DOI: 10.1137/110830289.

[21] M. Bulíček and J. Málek. “On unsteady internal flows of Bingham fluids subject to threshold slip on the impermeable boundary”. In: *Recent developments of mathematical fluid mechanics*. Ed. by H. Amann, Y. Giga, H. Kozono, H. Okamoto, and M. Yamazaki. Adv. Math. Fluid Mech. Birkhäuser/Springer, Basel, 2016, pp. 135–156.

[22] M. Bulíček, J. Málek, K. Rajagopal, and E. Süli. “On elastic solids with limiting small strain: modelling and analysis”. In: *EMS Surv. Math. Sci.* 1.2 (2014), pp. 283–332. DOI: 10.4171/EMSS/7.

[23] C. Carstensen and S. A. Funken. “Fully reliable localized error control in the FEM”. In: *SIAM J. Sci. Comput.* 21.4 (2000), pp. 1465–1484. DOI: 10.1137/S1064827597327486.

[24] C. Carstensen and R. Klose. “A posteriori finite element error control for the *p*-Laplace problem”. In: *SIAM J. Sci. Comput.* 25.3 (2003), pp. 792–814. DOI: 10.1137/S1064827502416617.

[25] A. Chaillou and M. Suri. “A posteriori estimation of the linearization error for strongly monotone nonlinear operators”. In: *J. Comput. Appl. Math.* 205.1 (2007), pp. 72–87. DOI: 10.1016/j.cam.2006.04.041.

[26] S.-K. Chua and R. L. Wheeden. “Estimates of best constants for weighted Poincaré inequalities on convex domains”. In: *Proc. London Math. Soc. (3)* 93.1 (2006), pp. 197–226. DOI: 10.1017/S0024611506015826.

[27] P. Ciarlet Jr. and M. Vohralík. “Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients”. In: *ESAIM Math. Model. Numer. Anal.* 52.5 (2018), pp. 2037–2064. DOI: 10.1051/m2an/2018034.

[28] A. Cohen, R. DeVore, and R. H. Nochetto. “Convergence rates of AFEM with H^{-1} data”. In: *Found. Comput. Math.* 12.5 (2012), pp. 671–718. DOI: 10.1007/s10208-012-9120-1.

[29] L. Diening and C. Kreuzer. “Linear convergence of an adaptive finite element method for the *p*-Laplacian equation”. In: *SIAM J. Numer. Anal.* 46.2 (2008), pp. 614–638.

[30] L. Diening, C. Kreuzer, and E. Süli. “Finite element approximation of steady flows of incompressible fluids with implicit power-law-like rheology”. In: *SIAM J. Numer. Anal.* 51.2 (2013), pp. 984–1015. DOI: 10.1137/120873133.

[31] L. El Alaoui, A. Ern, and M. Vohralík. “Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems”. In: *Comput. Methods Appl. Mech. Engrg.* 200.37-40 (2011), pp. 2782–2795. DOI: 10.1016/j.cma.2010.03.024.

[32] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*. Vol. 159. Applied Mathematical Sciences. New York: Springer-Verlag, 2004, pp. xiv+524.

[33] A. Ern and M. Vohralík. “Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs”. In: *SIAM J. Sci. Comput.* 35.4 (2013), A1761–A1791. DOI: 10.1137/120896918.

[34] A. Ern and M. Vohralík. “Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations”. In: *SIAM J. Numer. Anal.* 53.2 (2015), pp. 1058–1081. DOI: 10.1137/130950100.

[35] L. C. Evans. *Partial differential equations.* Vol. 19. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 1998, pp. xviii+662.

[36] G. Francfort, F. Murat, and L. Tartar. “Monotone operators in divergence form with x-dependent multivalued graphs”. In: *Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8)* 7.1 (2004), pp. 23–59.

[37] M. Griebel and M. A. Schweitzer, eds. *Meshfree methods for partial differential equations VIII.* Vol. 115. Lecture Notes in Computational Science and Engineering. Selected papers from the 8th International Workshop held in Bonn, September 7–9, 2015. Springer, Cham, 2017, pp. viii+231.

[38] J. Hron, J. Málek, J. Stebel, and K. Touška. “A novel view on computations of steady flows of Bingham fluids using implicit constitutive relations”. MORE preprint MORE/2017/08, submitted for publication. 2017. URL: `http://ncmm.karlin.mff.cuni.cz/db/attachments/single/417`.

[39] P. Jiránek, Z. Strakoš, and M. Vohralík. “A posteriori error estimates including algebraic error and stopping criteria for iterative solvers”. In: *SIAM J. Sci. Comput.* 32.3 (2010), pp. 1567–1590. DOI: `10.1137/08073706X`.

[40] C. Kreuzer and E. Süli. “Adaptive finite element approximation of steady flows of incompressible fluids with implicit power-law-like rheology”. In: *ESAIM Math. Model. Numer. Anal.* 50.5 (2016), pp. 1333–1369. DOI: `10.1051/m2an/2015085`.

[41] V. Kulvait, J. Málek, and K. Rajagopal. “Modeling Gum Metal and other newly developed titanium alloys within a new class of constitutive relations for elastic bodies.” In: *Archives of Mechanics* 69.3 (2017).

[42] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires.* Dunod; Gauthier-Villars, Paris, 1969, pp. xx+554.

[43] J. M. Melenk and I. Babuška. “The partition of unity finite element method: basic theory and applications”. In: *Comput. Methods Appl. Mech. Engrg.* 139.1-4 (1996), pp. 289–314. DOI: `10.1016/S0045-7825(96)01087-0`.

[44] J. Papež, U. Rüde, M. Vohralík, and B. Wohlmuth. “Sharp algebraic and total a posteriori error bounds for h and p finite elements via a multilevel approach”. HAL preprint 01662944, submitted for publication. 2017. URL: `https://hal.inria.fr/hal-01662944/`.

[45] J. Papež, Z. Strakoš, and M. Vohralík. “Estimating and localizing the algebraic and total numerical errors using flux reconstructions”. In: *Numer. Math.* 138.3 (2018), pp. 681–721. DOI: `10.1007/s00211-017-0915-5`.

[46] S. Repin. *A posteriori estimates for partial differential equations.* Vol. 4. Radon Series on Computational and Applied Mathematics. Walter de Gruyter GmbH & Co. KG, Berlin, 2008, pp. xii+316. DOI: `10.1515/9783110203042`.

[47] G. Strang and G. J. Fix. *An analysis of the finite element method.* Vol. 212. Prentice-hall Englewood Cliffs, NJ, 1973.

[48] H. Triebel. *Function spaces and wavelets on domains.* Vol. 7. EMS Tracts in Mathematics. European Mathematical Society (EMS), Zürich, 2008, pp. x+256. DOI: `10.4171/019`.

[49] A. Veeser. “Approximating gradients with continuous piecewise polynomial functions”. In: *Found. Comput. Math.* 16.3 (2016), pp. 723–750. DOI: `10.1007/s10208-015-9262-z`.

[50] A. Veeser and R. Verfürth. “Explicit upper bounds for dual norms of residuals”. In: *SIAM J. Numer. Anal.* 47.3 (2009), pp. 2387–2405. DOI: `10.1137/080738283`.

[51] A. Veeser and R. Verfürth. “Poincaré constants for finite element stars”. In: *IMA J. Numer. Anal.* 32.1 (2012), pp. 30–47. DOI: `10.1093/imanum/drr011`.

[52] R. Verfürth. “A posteriori error estimates for non-linear parabolic equations”. Tech. report, Ruhr-Universität Bochum. 2004.

[53] R. Verfürth. *A posteriori error estimation techniques for finite element methods.* Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press, 2013, pp. xx+393. DOI: `10.1093/acprof:oso/9780199679423.001.0001`.

Chapter III

Analysis of PCD preconditioner for Navier-Stokes equations

> It is a tool that does suck up dust to make what you walk on in a home tidy.
>
> Berchenko-Kogan [4]

1 Introduction

In this chapter we are concerned with a boundary-value problem for steady flows of Navier-Stokes fluid. Given a domain $\Omega \subset \mathbb{R}^3$, an inflow boundary $\Gamma_{\text{in}} \subset \partial\Omega$, an outflow boundary $\Gamma_{\text{out}} \subset \partial\Omega$, a body volumetric force f, and an inflow velocity $\mathbf{v}^{\text{D}}$, the problem is to find velocity $\mathbf{v}$ and pressure p such that

$$\mathbf{v}\cdot\nabla\mathbf{v} - \Delta\mathbf{v} - \nabla p = \mathbf{f} \qquad \text{in } \Omega, \tag{1.1a}$$

$$\operatorname{div}\mathbf{v} = 0 \qquad \text{in } \Omega, \tag{1.1b}$$

$$\mathbf{v} = \mathbf{v}^{\text{D}} \qquad \text{on } \Gamma_{\text{in}}, \tag{1.1c}$$

$$\mathbf{v} = \mathbf{0} \qquad \text{on } \partial\Omega\backslash\left(\Gamma_{\text{in}} \cup \Gamma_{\text{out}}\right), \tag{1.1d}$$

$$\frac{\partial\mathbf{v}}{\partial\mathbf{n}} - p\mathbf{n} = \mathbf{0} \qquad \text{on } \Gamma_{\text{out}}, \tag{1.1e}$$

where $\mathbf{n}$ is the unit outer normal to $\partial\Omega$.

Note that despite the importance of problem (1.1) in science and engineering, this is a difficult problem from the standpoint of mathematical analysis. The classical theory of weak solutions for incompressible flows by Leray, Hopf, Schauder, Ladyzhenskaya, Lions, Solonnikov, and others (see [37] and references therein) is not directly applicable to this problem because the *do-nothing* boundary condition (1.1e) prevents validity of a priori estimates. Roughly said, it is difficult, if not impossible, to a priori guarantee the outflow boundary Γ_{out} to carry away a sufficient amount of kinetic energy (which is brought from the inflow Γ_{in} or created by the volumetric force $\mathbf{f}$). A number of remedies, mostly based on certain modifications of (1.1e), have been proposed; see, e.g., [54, 7], also [38]. We ignore this issue as it is not important for the subsequent exposition. However, it should be remembered that it is hard to a priori guarantee the sign of $\mathbf{v}\cdot\mathbf{n}$ on Γ_{out} for a solution of (1.1), if it exists, as well as for a sequence of numerical approximations. On the other hand, the direction of the Dirichlet datum on Γ_{in} is in many applications inwards so that $\mathbf{v}\cdot\mathbf{n} = \mathbf{v}^{\text{D}}\cdot\mathbf{n} \leq 0$ on Γ_{in}, both for a solution of (1.1), if it exists, and properly constructed numerical approximations.

We consider a linearization of system (1.1), namely Oseen linearization which is equivalent to the Picard iteration. The linear boundary-value problem to be solved during one nonlinear iteration is then to find a velocity $\mathbf{v}$ and a pressure p for a fixed wind $\mathbf{b}$, a parameter $\alpha \in [0, 1]$, and the other previously fixed data, such that

$$
\begin{aligned}
(1-\alpha)\mathbf{b}\cdot\nabla\mathbf{v}+\alpha\operatorname{div}(\mathbf{v}\otimes\mathbf{b})-\Delta\mathbf{v}-\nabla p&=\mathbf{f} && \text{in } \Omega, && (1.2\text{a})\\
\operatorname{div}\mathbf{v}&=0 && \text{in } \Omega, && (1.2\text{b})\\
\mathbf{v}&=\mathbf{v}^{\mathrm{D}} && \text{on } \Gamma_{\text{in}}, && (1.2\text{c})\\
\mathbf{v}&=\mathbf{0} && \text{on } \partial\Omega\setminus\left(\Gamma_{\text{in}}\cup\Gamma_{\text{out}}\right), && (1.2\text{d})\\
\frac{\partial\mathbf{v}}{\partial\mathbf{n}}-p\mathbf{n}&=\mathbf{0} && \text{on } \Gamma_{\text{out}}. && (1.2\text{e})
\end{aligned}
$$

In the subsequent analysis, we may need to assume either that

$$\mathbf{b}\cdot\mathbf{n}\le 0 \qquad \text{on } \Gamma_{\text{in}}, \tag{1.3}$$

or

$$\mathbf{b}\cdot\mathbf{n}\ge 0 \qquad \text{on } \Gamma_{\text{out}}, \tag{1.4}$$

depending on the particular choice of preconditioner. We keep in mind, in the spirit of the preceding paragraph, that (1.3) is, for many numerical schemes, much less restrictive than (1.4). The effort to obtain a priori estimates for (1.1) brings another lesson: testing by the solution makes the convective term disappear (up to the aforementioned issue of uncontrollable energy flux through the outflow boundary Γ_{out}) thanks to the property (1.1b) ensuring conservative energy transport; mimicking energy estimates for the linearized problem (1.2) may require the assumption that $\operatorname{div}\mathbf{b}$ is in a certain sense small, so that the linearized kinetic energy transport is close to conservative. First observe that if $\operatorname{div}\mathbf{b}$ vanished pointwise, the problem (1.2) would be equivalent for all $\alpha\in[0,1]$. The special distinct cases are $\alpha=0$ (convective form), $\alpha=\frac{1}{2}$ (skew-symmetric form), and $\alpha=1$ (conservative form).[1] With the skew-symmetric form the smallness of $\operatorname{div}\mathbf{b}$ is not needed to obtain a priori estimates because testing by a solution makes the whole convective term disappear.

We would like to point out that a Newton linearization of (1.1) is not in principle excluded; more precisely, the following analysis of the preconditioner does not rely on a particular linearization; (1.2) is considered a model problem which under mild assumptions admits energy estimates. In fact, additional terms, should they appear in (1.2) from a derivative of (1.1), easily render the convective term linearization too far from the skew-symmetry unless $\mathbf{b}$ is close to a solution.

For the sake of construction of the preconditioner we write problem (1.2) in the block form

$$Q\begin{pmatrix}\mathbf{v}\\ p\end{pmatrix}=\begin{pmatrix}\mathbf{f}\\ 0\end{pmatrix} \tag{1.5}$$

with operator Q (acting from a Cartesian product of a suitable velocity space and pressure space to its dual) given by

$$Q=\begin{pmatrix}-\Delta+(1-\alpha)\mathbf{b}\cdot\nabla+\alpha\operatorname{div}(\bullet\otimes\mathbf{b}) & \nabla\\ -\operatorname{div} & 0\end{pmatrix}. \tag{1.6}$$

A good approximation to Q is an upper triangular approximation of its Schur complement factorization

$$P_{\mp}:=\begin{pmatrix}-\Delta+(1-\alpha)\mathbf{b}\cdot\nabla+\alpha\operatorname{div}(\bullet\otimes\mathbf{b}) & \nabla\\ 0 & \mp S\end{pmatrix}, \tag{1.7}$$

where the pressure Schur complement S is

$$S=-\operatorname{div}\left(-\Delta+(1-\alpha)\mathbf{b}\cdot\nabla+\alpha\operatorname{div}(\bullet\otimes\mathbf{b})\right)^{-1}\nabla. \tag{1.8}$$

[1] The skew-symmetric form is preferred for convergence analysis of velocity-pressure numerical schemes as testing by a solution makes the skew-symmetric term disappear (even when the discrete velocity is not divergence-free) in order to recover energy inequality, see, e.g., [60, 16]. The conservative form is of special interest for non-Newtonian fluids in $W^{1,r}$ spaces with small r but it seems to only work with pointwise divergence-free velocity approximations, see [16, section 3.2].

The preconditioned operator then reads

$$QP_{\mp}^{-1} = \begin{pmatrix} I & 0 \\ -\operatorname{div}\left(-\Delta + (1-\alpha)\mathbf{b}\cdot\nabla + \alpha\operatorname{div}\left(\bullet\otimes\mathbf{b}\right)\right)^{-1} & \pm I \end{pmatrix}. \tag{1.9}$$

The GMRES method applied to $QP_{\pm}^{-1}$ converges in at most two iterations because the minimal polynomial of $QP_{\pm}^{-1}$ is $p(t) = (t-1)(t\mp 1)$, which is the argument due to Murphy, Golub, and Wathen [49]. But the Schur complement S is non-local for any local discretization, e.g., FEM, of (1.5), and hence its inversion is not computationally feasible. A possible remedy is to further approximate S by formally swapping the order of operators in (1.8). This results in a pair of possible approximations to the inverse of the Schur complement S^{-1}:

$$S^{-1} \approx \left(-\Delta + (1-\alpha)\mathbf{b}\cdot\nabla + \alpha\operatorname{div}\left(\mathbf{b}\bullet\right)\right)(-\Delta)^{-1} =: X^{-1} \tag{1.10}$$

and

$$S^{-1} \approx (-\Delta)^{-1}\left(-\Delta + (1-\alpha)\mathbf{b}\cdot\nabla + \alpha\operatorname{div}\left(\mathbf{b}\bullet\right)\right) =: Y^{-1}. \tag{1.11}$$

This idea first appeared in studies [33, 57, 34]. It was originally motivated by constructing an approximation to the Green's function of the Schur complement solution operator S^{-1} by approximating the Green's function of the Laplace operator by the Poisson fundamental solution. This corresponds to the fact that, provided the wind $\mathbf{b}$ is constant, the commutation leading to (1.10) or (1.11) is exact unless a boundary is present, i.e., the commutation holds if periodic boundary conditions are assumed or $\Omega = \mathbb{R}^3$. In this case, $X^{-1} = Y^{-1} = S^{-1}$. In the case of the boundary-value problem (1.2) the commutation is not exact and applicability of the approach strongly depends on the boundary conditions used to define the particular operators in (1.10) or (1.11). This is by far not an obvious problem with the exception of enclosed flows, i.e., $\Gamma_{\text{in}} = \Gamma_{\text{out}} = \emptyset$, where it is more or less obvious that natural boundary conditions are appropriate.

This approach has been named the *pressure convection-diffusion* (PCD) preconditioner due to the presence of the convection-diffusion operator $(-\Delta + (1-\alpha)\mathbf{b}\cdot\nabla + \alpha\operatorname{div}(\mathbf{b}\bullet))$ in (1.10). The variant (1.11) has been considered later [23, 22]. It has been immediately noticed that PCD can be viewed as an extension of the "Stokes" preconditioner $S^{-1} \approx I$. Indeed, when $\mathbf{b} = \mathbf{0}$, both PCD variants take the form $X^{-1} = Y^{-1} = I$, which is a good preconditioner for the Schur complement of the Stokes problem and also for small data (low Reynolds number, or high viscosity) Navier-Stokes problem. See [19, p. 1300] for a comprehensive list of references; of particular note is the field-of-value analysis by Klawonn and Starke [35], who show that the preconditioned operator QP_{-}^{-1}, with the approximation $S \approx I$ in (1.7), has a numerical range bounded away from zero and infinity uniformly in discretization. Nevertheless, the dependence on data size (Reynolds number) is severe and this preconditioner quickly becomes ineffective with increasing data. Thus the compact perturbation proportional to $\mathbf{b}$ in (1.10), (1.11) can be viewed to be balancing the compact perturbation in S growing with $\mathbf{b}$. This is actually an important philosophical point of the analysis provided in this work. Rather then trying to evaluate deviation from the commutation, i.e., the smallness of $SX^{-1} - I$ or $SY^{-1} - I$, which would be zero if the commutation was exact, we are concerned with the smallness of $SX^{-1} - S^{\infty}$, $SY^{-1} - S^{\infty}$, respectively, where $S^{\infty} = -\operatorname{div}(-\Delta)^{-1}\nabla$ is the Stokes Schur complement. This is motivated by the following. While $SX^{-1} - I$, $SY^{-1} - I$ vanish in the no-boundary situation (and constant wind $\mathbf{b}$), they are certainly non-zero as long as a boundary is present. On the other hand $SX^{-1} - S^{\infty}$, $SY^{-1} - S^{\infty}$ vanish whenever $\mathbf{b} = \mathbf{0}$ even in the presence of the boundary. Hence the smallness of $SX^{-1} - S^{\infty}$, $SY^{-1} - S^{\infty}$ expresses the ability of the preconditioner to compensate for the departure from the Stokes case $\mathbf{b} = \mathbf{0}$ and corresponds to the expected deterioration of the preconditioner with increasing $\mathbf{b}$. We remark that any published a priori estimates for SX^{-1}, SY^{-1} were using the worst-case estimate $\|SX^{-1}\| \leq \|S\|\|X^{-1}\|$, and similarly with Y^{-1}. Any such estimate is worse than an analogous estimate for the Stokes preconditioner $S^{-1} \approx I$, when the preconditioned Schur complement is just S, and such estimate is thus not able to explain the success of the PCD correction $X^{-1} - I$, $Y^{-1} - I$, which compensates for the perturbation $S - S^{\infty}$. The important observation thus is that the improvement of PCD over the Stokes preconditioner has never been quantitatively explained. To this point, we observe that $SX^{-1} - S^\infty$ is compact and can be written as a formal commutator of certain three differential operators; see (2.54). This makes it possibly amenable to Fourier analysis, which we were not fully successful with but which we sketch in Section 2.7. We stress once more that the synergistic effect of the compact perturbations $S - S^\infty$ and $X^{-1} - I$, which are both proportional to $\mathbf{b}$, still remains to be fully explained. The same holds mutatis mutandis for $SY^{-1} - S^\infty$; see (2.76).

In Section 2 we develop a theory for the PCD preconditioner in the setting of infinite-dimensional function spaces.[2] The setting allows us to treat both variants of the PCD preconditioner and we obtain artificial boundary conditions needed for both versions as a consequence of the effort to obtain a priori estimates and invertibility of the preconditioner. A distinctive feature of our analysis is that we do not require restrictive conditions on the wind $\mathbf{b}$ such as uniform bounds on $\|\mathbf{b}\|_\infty$ and $\|\operatorname{div}\mathbf{b}\|_3$, or even the assumption $\operatorname{div}\mathbf{b} = 0$, which appear more or less explicitly in existing studies. Specifically, we only require the aforementioned correct direction of the wind on the part of the boundary, i.e., the sign of $\mathbf{b}\cdot\mathbf{n}$, the smallness of $\operatorname{div}\mathbf{b}$ in a certain sense, and a uniform bound in a native energy space $\|\mathbf{b}\|_{1,2}$. The lack of a need for extra regularity of wind is balanced in our study by an assumption of the $W^{1,3+\epsilon}$-regularity for the Dirichlet-Neumann Laplacian problem. It turns out that existing literature provides such estimates under very reasonable conditions which are typically met in practice; see (2.2). The section continues with an analysis of the GMRES method; specifically we relate the convergence of the GMRES method applied to the preconditioned system to the convergence of the GMRES method applied to the preconditioned Schur complement. That allows us to simplify the analysis by considering only SX^{-1}, SY^{-1}. Important structural observations about the preconditioned Schur complement are made and a certain worse-case estimate for the convergence rate is obtained. We make progress towards the goal of explaining the approximation quality of the PCD operator to the Schur complement. We sketch a simplified analysis of the synergistic effect of the PCD correction in the preconditioner and the convection perturbation in the Schur complement at the end of the section.

Section 3 provides a precise methodology for the construction of a discrete PCD operator, first in a general setting, and subsequently applied to some important specific discretizations. We obtain several novel variants of PCD, but also one which has been previously described by a verbal description, which is, in our opinion, prone to possible misinterpretation. The constructed operators are, under appropriate conditions, guaranteed to be invertible and inherit in certain cases the a priori bounds of Section 2. We make progress towards establishing approximation and convergence properties of the PCD preconditioner by transfering some of the results of Section 2 derived for infinite-dimensional operators to the derived discrete cases. For that there are missing pieces which might be quite technical and difficult to obtain, e.g., validity of the aforementioned $W^{1,3+\epsilon}$-regularity (2.2) in the discrete case. We close the section by commenting on already published works on PCD and comparing with our results.

Appendix A collects some results of functional analysis which we need in Section 2 and in Appendix B. Appendix B contains new results concerning convergence of the GMRES method under compact perturbations, which were obtained as a byproduct of this research and are used in Section 2.

2 Analysis of PCD in infinite-dimensional spaces

2.1 Preliminaries

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded Lipschitz domain and Γ, D subdomains of $\partial\Omega$ of positive measure. Let $\mathbf{n}$ denote the outer unit normal of $\partial\Omega$. We will fix the definition of Γ and D later. We denote the Lebesgue norm $\|u\|_p := \left(\int_\Omega |u|^p\right)^{\frac{1}{p}}$ for any measurable function u and similarly $\|\mathbf{u}\|_p := \left(\int_\Omega |\mathbf{u}|^p\right)^{\frac{1}{p}}$. We define the usual Sobolev space $\mathbf{W}^{1,p}$ consisting of all measurable

[2] This approach is often called *operator preconditioning* but one could call that just analysis. For examples of applications of the approach see the survey monograph by Málek and Strakoš [45] and the references therein. The advantage of the approach is that the analysis does not rely on any assumptions about discretization and its properties, which are often very technical. That allows to establish desirable properties of the preconditioner in the PDE context and eventually transfer the properties to its discretized version.

vector-valued functions $\mathbf{u} : \Omega \to \mathbb{R}^d$ having finite norm $\|\mathbf{u}\|_{1,p} \coloneqq \left(\int_\Omega \left(|\mathbf{u}|^2 + |\nabla \mathbf{u}|^2 \right)^{\frac{p}{2}} \right)^{\frac{1}{p}}$, and define the usual Sobolev space $W_\Gamma^{1,p}$ consisting of measurable scalar functions $u : \Omega \to \mathbb{R}$ with zero trace on Γ having finite norm $\|\nabla u\|_p \coloneqq \left(\int_\Omega |\nabla u|^p \right)^{\frac{1}{p}}$, and similarly the vector-valued space $\mathbf{W}_\mathrm{D}^{1,2}$ of functions vanishing on D with the norm $\|\nabla \mathbf{u}\|_2$. We define the Sobolov-Poincaré embedding constant

$$C_\mathrm{P}(p, \Omega, \Gamma) \coloneqq \sup_{r \in W_\Gamma^{1,p}} \frac{\|r\|_{p^*}}{\|\nabla r\|_p}$$

where $p^* = \frac{3p}{3-p}$ when $1 \le p < 3$ and $p^* = \infty$ when $p > 3$.

Now we can fix the definition of the Laplacian solve in the PCD operator (1.10) or (1.11). We denote by $A_\Gamma : W_\Gamma^{1,2} \to (W_\Gamma^{1,2})^\#$ a Laplace operator restricted to $W_\Gamma^{1,2}$, i.e., $\langle A_\Gamma r, q \rangle \coloneqq \int_\Omega \nabla r \cdot \nabla q$ for any $r, q \in W_\Gamma^{1,2}$. By the standard theory $A_\Gamma^{-1} \in \mathcal{L}((W_\Gamma^{1,2})^\#, W_\Gamma^{1,2})$ and it is a solution operator for the mixed Poisson problem

$$-\Delta r = s \text{ in } \Omega, \tag{2.1a}$$

$$r = 0 \text{ on } \Gamma, \tag{2.1b}$$

$$\frac{\partial r}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega \backslash \Gamma, \tag{2.1c}$$

i.e., we write $r = A_\Gamma^{-1} s$.

We will also use a further regularity assumption on A_Γ^{-1}. If the Lipschitz domain Ω is additionally a *creased* Lipschitz domain then there exists $\epsilon(\Omega, \Gamma) > 0$ and $C(\Omega, \Gamma) > 0$ such that

$$\|\nabla r\|_{3+\epsilon(\Omega,\Gamma)} \le C(\Omega, \Gamma) \|s\|_{\left(W_\Gamma^{1, \frac{3+\epsilon(\Omega,\Gamma)}{2+\epsilon(\Omega,\Gamma)}} \right)^\#} \tag{2.2}$$

whenever $r = A_\Gamma^{-1} s$, see [47, Theorem 8.2, Corollary 8.3]. Roughly said, a creased Lipschitz domain is a Lipschitz domain with interior angle less than π at the intersection of the Dirichlet boundary Γ and the Neumann boundary $\partial\Omega \backslash \Gamma$, thus avoiding the critical Zaremba singularity. See [47, section 2, Definition 2.3] for a precise definition.

Summarizing the last two paragraphs, we have that $A_\Gamma^{-1} \in \mathcal{L}\left(L^2, W_\Gamma^{1,2}\right)$ and A_Γ^{-1} is even compact in this topology thanks to the Rellich-Kondrachov embedding. Furthermore, when Ω is a creased Lipschitz domain, then A_Γ^{-1} is compact in $\mathcal{L}\left(L^2, W_\Gamma^{1,3}\right)$ and also in $\mathcal{L}\left(L^2, W_\Gamma^{1,3+\epsilon(\Omega,\Gamma)}\right)$ for some $\epsilon(\Omega, \Gamma) > 0$. Let us point out that the norm of A_Γ^{-1} in all these topologies depends solely on Ω and Γ.

Next we fix the definition of the convection operator $K_{\alpha,\mathbf{w},\Gamma}$. For parameter $\alpha \in [0, 1]$ and wind $\mathbf{w} \in \mathbf{W}^{1,2}$ we define $K_{\alpha,\mathbf{w},\Gamma}$ by duality

$$\langle K_{\alpha,\mathbf{w},\Gamma} r, q \rangle = \int_\Omega \mathbf{w} \cdot \nabla r \, q + \alpha \operatorname{div} \mathbf{w} \, r \, q$$

$$\text{whenever } r, q \in W_\Gamma^{1,2} \text{ or } r \in L^2,\ q \in W_\Gamma^{1,3+\epsilon}, \quad \begin{cases} \epsilon \ge 0 & \text{if } \alpha = 0, \\ \epsilon > 0 & \text{if } \alpha \in (0, 1]. \end{cases} \tag{2.3}$$

This definition includes the standard convective term when $\alpha = 0$, the skew-symmetric form of the convective term when $\alpha = \frac{1}{2}$, and the conservative form when $\alpha = 1$. Obviously $K_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(W_\Gamma^{1,2}, (W_\Gamma^{1,2})^\#)$ thanks to the Sobolev embedding $W^{1,2} \hookrightarrow L^6$ and also $K_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(W_\Gamma^{1,3+\epsilon}, L^2)$ with any ϵ from (2.3).

Now we are in the position to characterize the injectivity of the convection-diffusion operator $F_{\alpha,\mathbf{w},\Gamma} \coloneqq A_\Gamma + K_{\alpha,\mathbf{w},\Gamma}$, which is also bounded in $\mathcal{L}(W_\Gamma^{1,2}, (W_\Gamma^{1,2})^\#)$ for any $\mathbf{w} \in \mathbf{W}^{1,2}$.

Lemma 2.1 (Ellipticity of $F_{\alpha,\mathbf{w},\Gamma}$)**.** *Let* $\mathbf{w} \in \mathbf{W}^{1,2}$ *and* $\mathbf{w} \cdot \mathbf{n} \ge 0$ *on* $\partial\Omega \backslash \Gamma$*. Then it holds that*

$$\inf_{r \in W_\Gamma^{1,2}} \frac{\langle F_{\alpha,\mathbf{w},\Gamma} r, r \rangle}{\|\nabla r\|_2^2} \ge 1 - \left| \alpha - \tfrac{1}{2} \right| C_\mathrm{P}(2, \Omega, \Gamma)^2 \|\operatorname{div} \mathbf{w}\|_{\frac{3}{2}}. \tag{2.4}$$

Proof. Fix $r \in W^{1,2}_\Gamma$ and estimate using integration by parts (noticing $r = 0$ on Γ), the nonnegativity of $\mathbf{w} \cdot \mathbf{n}$ on $\partial\Omega\backslash\Gamma$, Hölder's inequality, and the Sobolev-Poincaré inequality $\|r\|_6 \leq C_\mathrm{P}(2, \Omega, \Gamma)\|\nabla r\|_2$:

$$\begin{aligned}\langle F_{\alpha,\mathbf{w},\Gamma} r, r\rangle &= \|\nabla r\|_2^2 + \int_\Omega \mathbf{w} \cdot \nabla \frac{r^2}{2} + 2\alpha \int_\Omega \operatorname{div} \mathbf{w} \frac{r^2}{2} \\ &= \|\nabla r\|_2^2 + (2\alpha - 1) \int_\Omega \operatorname{div} \mathbf{w} \frac{r^2}{2} + \int_{\partial\Omega\backslash\Gamma} \mathbf{w} \cdot \mathbf{n} \frac{r^2}{2} \\ &\geq \|\nabla r\|_2^2 - \frac{|2\alpha - 1|}{2} \| \operatorname{div} \mathbf{w}\|_{\frac{3}{2}} \|r\|_6^2 \\ &\geq \left(1 - \left|\alpha - \tfrac{1}{2}\right| C_\mathrm{P}(2, \Omega, \Gamma)^2 \| \operatorname{div} \mathbf{w}\|_{\frac{3}{2}}\right) \|\nabla r\|_2^2\end{aligned}$$

which is the desired estimate. □

Finally we can define the linear operator

$$X^{-1}_{\alpha,\mathbf{w},\Gamma} := F_{\alpha,\mathbf{w},\Gamma} A^{-1}_\Gamma = I + K_{\alpha,\mathbf{w},\Gamma} A^{-1}_\Gamma. \tag{2.5}$$

By the considerations above, the operator $X^{-1}_{\alpha,\mathbf{w},\Gamma}$ is bounded in $\mathcal{L}((W^{1,2}_\Gamma)^\#) := \mathcal{L}((W^{1,2}_\Gamma)^\#, (W^{1,2}_\Gamma)^\#)$ whenever $\Omega \subset \mathbb{R}^3$ is a Lipschitz domain, $\Gamma \subset \partial\Omega$ is open, $\alpha \in [0, 1]$, and $\mathbf{w} \in \mathbf{W}^{1,2}$. Note that this abstract setting will allow us to analyze both PCD versions (1.10) and (1.11) by considering either $X^{-1}_{\alpha,\mathbf{w},\Gamma}$ or $Y^{-1}_{\alpha,\mathbf{w},\Gamma} := X^{-\#}_{1-\alpha,-\mathbf{w},\Gamma}$, the adjoint of $X^{-1}_{1-\alpha,-\mathbf{w},\Gamma}$. We will discuss the construction of the Y-variant and the choice of Γ for both versions in detail in Section 2.6.

We will now assume general conditions we expect from the velocity convection-diffusion operator, which will be needed for subsequent analysis. Its example appeared informally in the upper-leftmost block of the operator in system (1.6). Let us assume that we are looking for a Picard velocity update in function space $\mathbf{W}^{1,2}_\mathrm{D} := (W^{1,2}_\mathrm{D})^3$ and for pressure in L^2. Here D is an open subdomain of $\partial\Omega$ of positive measure, representing the boundary conditions (1.2c), (1.2d). Indeed, since the Oseen system (1.2) is linear, we can subtract a previous Picard iterate assumed to fulfill boundary conditions (1.2c), (1.2d) and look for an update $\delta\mathbf{v}$, which fulfills $\delta\mathbf{v} = \mathbf{0}$ on D.

We assume that there is an operator $\mathbf{F} \in \mathcal{L}(\mathbf{W}^{1,2}_\mathrm{D}, (\mathbf{W}^{1,2}_\mathrm{D})^\#)$ such that

$$\underline{\mathbf{F}} := \inf_{\mathbf{v}\in\mathbf{W}^{1,2}_\mathrm{D}} \frac{\langle \mathbf{F}\mathbf{v}, \mathbf{v}\rangle}{\|\nabla \mathbf{v}\|_2^2} > 0. \tag{2.6}$$

By virtue of the Lax-Milgram theorem this implies $\mathbf{F}^{-1} \in \mathcal{L}((\mathbf{W}^{1,2}_\mathrm{D})^\#, \mathbf{W}^{1,2}_\mathrm{D})$ and $\|\mathbf{F}^{-1}\|_{\mathcal{L}((\mathbf{W}^{1,2}_\mathrm{D})^\#,\mathbf{W}^{1,2}_\mathrm{D})} \leq \underline{\mathbf{F}}^{-1}$.

If the operator $\mathbf{F}$ has similar structural properties as $F_{\alpha,\mathbf{w},\Gamma}$ then estimate (2.6) can be assured in a similar fashion as in Lemma 2.1. On the other hand if $\mathbf{F}$ comes from the Newton linearization of the term $\mathbf{v} \mapsto (1-\alpha)\mathbf{v}\cdot\nabla\mathbf{v}+\alpha \operatorname{div}(\mathbf{v}\otimes\mathbf{v})-\Delta\mathbf{v}$, then the ellipticity condition (2.6) might not hold unless $\mathbf{b}$ is close to a solution. In any case we will assume that (2.6) holds. Furthermore, we will assume that the convection part of $\mathbf{F}$ is compact in $\mathcal{L}(\mathbf{W}^{1,2}_\mathrm{D}, (\mathbf{W}^{1,2}_\mathrm{D})^\#)$, i.e., with definition of the velocity Laplacian $\mathbf{A} \in \mathcal{L}(\mathbf{W}^{1,2}_\mathrm{D}, (\mathbf{W}^{1,2}_\mathrm{D})^\#)$ by

$$\langle \mathbf{A}\boldsymbol{\phi}, \boldsymbol{\psi}\rangle = \int_\Omega \nabla\boldsymbol{\phi} : \nabla\boldsymbol{\psi} \qquad \text{for all } \boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbf{W}^{1,2}_\mathrm{D},$$

we assume that $\mathbf{K} := \mathbf{F} - \mathbf{A}$ is compact in $\mathcal{L}(\mathbf{W}^{1,2}_\mathrm{D}, (\mathbf{W}^{1,2}_\mathrm{D})^\#)$. This is true for both Picard and Newton linearization. But one can think of a more general situation, e.g., $\mathbf{F}$ featuring the SUPG stabilization, etc.; in such a case one has to verify the assumption on a case-by-case basis.

Now we can define the pressure Schur complement as a mapping given by $S := -\operatorname{div} \mathbf{F}^{-1}\nabla$, where $\operatorname{div} \in \mathcal{L}(\mathbf{W}^{1,2}_\mathrm{D}, L^2)$ as usual and $\nabla = -\operatorname{div}^\# \in \mathcal{L}(L^2, (\mathbf{W}^{1,2}_\mathrm{D})^\#)$. More precisely, $S : L^2 \to L^2$ is defined by $S : q \mapsto \operatorname{div} \boldsymbol{\phi}$ such that $\boldsymbol{\phi} \in \mathbf{W}^{1,2}_\mathrm{D}$ and

$$\langle \mathbf{F}\boldsymbol{\phi}, \boldsymbol{\psi}\rangle = \int_\Omega q \operatorname{div} \boldsymbol{\psi} \qquad \text{for all } \boldsymbol{\psi} \in \mathbf{W}^{1,2}_\mathrm{D};$$

the definition clearly makes sense due to the ellipticity assumption (2.6) and ensures that $S \in \mathcal{L}(L^2)$. The inf-sup constant of the divergence, or the Babuška-Brezzi constant, $\beta(\Omega, \mathrm{D})$ is defined as the largest constant fulfilling

$$\beta(\Omega, \mathrm{D})\,\|q\|_2 \leq \sup_{\mathbf{v}\in\mathbf{W}^{1,2}_{\mathrm{D}}} \frac{\int_\Omega q \operatorname{div} \mathbf{v}}{\|\nabla \mathbf{v}\|_2} = \|\nabla q\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\#} \qquad \text{for all } q \in L^2. \tag{2.7}$$

It is well known, provided that Ω is a Lipschitz domain and $\mathrm{D} = \partial\Omega$, that there exists $\beta(\Omega, \partial\Omega) > 0$ such that (2.7) holds for all $q \in L^2$ with $\int_\Omega q = 0$; see Lemma I.A.7. We will show that $\beta(\Omega, \mathrm{D})$ in (2.7) is also positive in the case $|\partial\Omega \setminus \mathrm{D}| > 0$, now with any $q \in L^2$, with arbitrary value of $\int_\Omega q$.

Lemma 2.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and D be an open subset of $\partial\Omega$ such that $|\partial\Omega \setminus \mathrm{D}| > 0$. Then there exists $\beta(\Omega, \mathrm{D}) > 0$ such that (2.7) holds.*

Proof. The strategy of the proof is to extend $f \in L^2(\Omega)$ by $f' \in L^2(\Omega')$ with certain Ω' such that $\int_{\Omega'} f' = 0$, which allows us to use the classical Bogovskiĭ operator for the no-slip boundary condition.

Extend Ω by a bounded Lipchitz domain $\Omega' \subset \mathbb{R}^d$ such that $\Omega' \supset \Omega$ and $\partial\Omega \cap \partial\Omega' = D$. Hence $\Omega' \setminus \Omega$ has positive measure. We will show that there exists $C(\Omega, \Omega') > 0$ such that for every $f \in L^2(\Omega)$ there exists $f' \in L^2(\Omega')$ such that $f' = f$ in Ω, $\int_{\Omega'} f' = 0$, and

$$\|f'\|_{2,\Omega'} \leq C(\Omega, \Omega')\|f\|_{2,\Omega}. \tag{2.8}$$

Fix $f \in L^2$, set $f' = f$ in Ω, and set f' to be a constant equal to $-\int_\Omega f/|\Omega' \setminus \Omega|$ in $\Omega' \setminus \Omega$. Hence $f' \in L^2(\Omega')$ and $\int_{\Omega'} f' = 0$. It remains to show (2.8). Then we have

$$\|f'\|^2_{2,\Omega'\setminus\Omega} = \int_{\Omega'\setminus\Omega} \frac{|\int_\Omega f|^2}{|\Omega' \setminus \Omega|^2} \leq \frac{|\Omega|}{|\Omega' \setminus \Omega|}\|f\|^2_{2,\Omega},$$

so that (2.8) holds with $C(\Omega, \Omega') \leq \sqrt{1 + \frac{|\Omega|}{|\Omega'\setminus\Omega|}} < \infty$.

Consider the Bogovskiĭ operator on the extended domain $\mathcal{B}' : L^2(\Omega')/\mathbb{R} \to \mathbf{W}^{1,2}_{\partial\Omega'}(\Omega')$, which fulfills

$$\begin{aligned} \operatorname{div} \mathcal{B}'g = g \quad \text{in } \Omega', \quad \|\nabla\mathcal{B}'g\|_{2,\Omega'} &\leq C_{\mathcal{B}'}\|g\|_{2,\Omega'} \\ \text{for all } g \in L^2(\Omega') \text{ with } &\int_{\Omega'} g = 0. \end{aligned} \tag{2.9}$$

Such $\mathcal{B}'$ indeed exists; see Remark I.A.8. Now for $f \in L^2(\Omega)$ consider the extension $f' \in L^2(\Omega')$ from the previous paragraph and define $\mathcal{B} : L^2(\Omega) \to \mathbf{W}^{1,2}_{\mathrm{D}}(\Omega)$ by $\mathcal{B}f = (\mathcal{B}'f')_{|\Omega}$. Using (2.8) and (2.9) immediately yields

$$\begin{aligned} \operatorname{div} \mathcal{B}g = g \quad \text{in } \Omega, \quad \|\nabla\mathcal{B}g\|_{2,\Omega} &\leq C(\Omega, \Omega')\, C_{\mathcal{B}'}\|g\|_{2,\Omega} \\ &\text{for all } g \in L^2(\Omega). \end{aligned} \tag{2.10}$$

Hence for any $q \in L^2(\Omega)$ we have, by (2.10),

$$\sup_{\mathbf{v}\in\mathbf{W}^{1,2}_{\mathrm{D}}} \frac{\int_\Omega q \operatorname{div} \mathbf{v}}{\|\nabla\mathbf{v}\|_{2,\Omega}} \geq \frac{\int_\Omega q \operatorname{div} \mathcal{B}q}{\|\nabla\mathcal{B}q\|_{2,\Omega}} \geq \frac{\|q\|^2_{2,\Omega}}{\|\nabla\mathcal{B}q\|_{2,\Omega}} \geq \frac{\|q\|_{2,\Omega}}{C(\Omega, \Omega')\, C_{\mathcal{B}'}},$$

which confirms that $\beta(\Omega, \mathrm{D})$ in (2.7) is positive and the proof is finished. □

Denote the norm of $\operatorname{div} : \mathbf{W}^{1,2}_{\mathrm{D}} \to L^2$ as

$$\Sigma(\Omega, \mathrm{D}) := \sup_{\mathbf{v}\in\mathbf{W}^{1,2}_{\mathrm{D}}} \frac{\|\operatorname{div} \mathbf{v}\|_2}{\|\nabla\mathbf{v}\|_2}. \tag{2.11}$$

Elementary computation shows that $\Sigma(\Omega, \mathrm{D}) \leq \sqrt{d}$ where d is a spatial dimension, i.e., for $\Omega \subset \mathbb{R}^d$. On the other hand $\Sigma(\Omega, \mathrm{D}) = 1$ whenever $\mathrm{D} = \partial\Omega$; this follows from integration by parts and density of smooth functions compactly-supported in $\Omega \cup (\partial\Omega \setminus \Gamma)$; see [59, equations (7), (8)]. In the sequel we are mostly concerned with the situation when $\Omega \subset \mathbb{R}^3$ is a bounded Lipschitz domain, D is an open subset of $\partial\Omega$ such that $|\mathrm{D}| > 0$ and $|\partial\Omega \setminus \mathrm{D}| > 0$.[3] In this case Lemma 2.2 yields $\beta(\Omega, \mathrm{D}) > 0$ and $\|\nabla \cdot\|_2$ is a norm on $\mathbf{W}_\mathrm{D}^{1,2}$ so that $\|\operatorname{div}\|_{\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, L^2)} = \Sigma(\Omega, \mathrm{D}) \le \sqrt{3}$. We can continue with a priori estimates for the Schur complement.

Lemma 2.3. *Let the conditions of Lemma 2.2 be fulfilled. Furthermore assume that* $\mathbf{F} \in \mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$ *satisfies estimate* (2.6)*. Then* $S = -\operatorname{div} F_{\alpha,\mathbf{w},\Gamma}^{-1} \nabla$ *fulfills*

$$\|S\|_{\mathcal{L}(L^2)} \le \Sigma(\Omega, \mathrm{D})^2 \|\mathbf{F}^{-1}\|_{\mathcal{L}((\mathbf{W}_\mathrm{D}^{1,2})^\#, \mathbf{W}_\mathrm{D}^{1,2})} \le \Sigma(\Omega, \mathrm{D})^2 \underline{\mathbf{F}}^{-1}, \tag{2.12}$$

$$\underline{S} := \inf_{q \in L^2} \frac{\int_\Omega S q\, q}{\|q\|_2^2} \ge \frac{\beta(\Omega, \mathrm{D})^2\, \underline{\mathbf{F}}}{\|\mathbf{F}\|^2_{\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)}}, \tag{2.13}$$

$$\|S^{-1}\|_{\mathcal{L}(L^2)} \le \underline{S}^{-1} \le \frac{\|\mathbf{F}\|^2_{\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)}}{\beta(\Omega, \mathrm{D})^2\, \underline{\mathbf{F}}}. \tag{2.14}$$

Proof. The first inequality of (2.12) comes from the definition considering that $\|\nabla\|_{\mathcal{L}(L^2, (\mathbf{W}_\mathrm{D}^{1,2})^\#)} = \|\operatorname{div}\|_{\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, L^2)} = \Sigma(\Omega, \mathrm{D})$.

The second inequality of (2.12) is the standard Lax-Milgram theory, see (2.6) above. By the same argument, (2.14) is a consequence of (2.13).

The ellipticity estimate (2.13) is proved with the aid of (2.7) by the following chain of inequalities

$$\begin{aligned}
\inf_{q \in L^2} \frac{\int_\Omega S q\, q}{\|q\|_2^2} &= \inf_{q \in L^2} \frac{\langle \mathbf{F}^{-1} \nabla q, \nabla q \rangle}{\|q\|_2^2} \ge \beta(\Omega, \mathrm{D})^2 \inf_{q \in L^2} \frac{\langle \mathbf{F}^{-1} \nabla q, \nabla q \rangle}{\|\nabla q\|^2_{(\mathbf{W}_\mathrm{D}^{1,2})^\#}} \\
&= \beta(\Omega, \mathrm{D})^2 \inf_{\mathbf{z} \in \nabla L^2} \frac{\langle \mathbf{F}^{-1} \mathbf{z}, \mathbf{z} \rangle}{\|\mathbf{z}\|^2_{(\mathbf{W}_\mathrm{D}^{1,2})^\#}} \ge \beta(\Omega, \mathrm{D})^2 \inf_{\mathbf{z} \in (\mathbf{W}_\mathrm{D}^{1,2})^\#} \frac{\langle \mathbf{F}^{-1} \mathbf{z}, \mathbf{z} \rangle}{\|\mathbf{z}\|^2_{(\mathbf{W}_\mathrm{D}^{1,2})^\#}} \\
&= \beta(\Omega, \mathrm{D})^2 \inf_{\mathbf{v} \in \mathbf{W}_\mathrm{D}^{1,2}} \frac{\langle \mathbf{F}\mathbf{v}, \mathbf{v} \rangle}{\|\mathbf{F}\mathbf{v}\|^2_{(\mathbf{W}_\mathrm{D}^{1,2})^\#}} \ge \frac{\beta(\Omega, \mathrm{D})^2}{\|\mathbf{F}\|^2_{\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)}} \inf_{\mathbf{v} \in \mathbf{W}_\mathrm{D}^{1,2}} \frac{\langle \mathbf{F}\mathbf{v}, \mathbf{v} \rangle}{\|\nabla \mathbf{v}\|_2^2}.
\end{aligned}$$

□

Analogously we define the Stokes Schur complement $S^\infty = -\operatorname{div} \mathbf{A}^{-1} \nabla$, for which Lemma 2.3 simplifies to the following.

Lemma 2.4. *Let the conditions of Lemma 2.2 be fulfilled. Then*

$$\|S^\infty\|_{\mathcal{L}(L^2)} = \Sigma(\Omega, \mathrm{D})^2, \qquad \|(S^\infty)^{-1}\|^{-1}_{\mathcal{L}(L^2)} = \inf_{q \in L^2} \frac{\int_\Omega S^\infty q\, q}{\|q\|_2^2} = \beta(\Omega, \mathrm{D})^2. \tag{2.15}$$

We leave the proof as an exercise.

Study of the Stokes Schur complement S^∞, also called the Cosserat operator, dates back to work by brothers Cosserat and Cosserat [10]; see also the survey [36]. The recent work by Costabel et al. [12] provides a summary of existing results and certain new developments in the theory of essential spectrum of S^∞ for the case $\mathrm{D} = \partial\Omega$. We are not aware of a corresponding result for general D.[4] Nevertheless the moral is that $\sigma_{\mathrm{ess}}(S^\infty)$ on corner domains contains non-trivial intervals.

[3] The case $\mathrm{D} = \emptyset$ is not interesting for applications. The case $\mathrm{D} = \partial\Omega$ requires changing the pressure space L^2 into $L^2/\mathbb{R} \cong \{q \in L^2, \int_\Omega q = 0\}$ in (2.7) in order to get $\beta(\Omega, \partial\Omega) > 0$, but most of the following results stay true mutatis mutandis.

[4] Costabel et al. [12] show for Ω planar polygonal domain and $\mathrm{D} = \partial\Omega$ that every corner of opening ω contributes to the essential spectrum of S^∞ by the interval $[\frac{1}{2} - \frac{|\sin\omega|}{2\omega}, \frac{1}{2} + \frac{|\sin\omega|}{2\omega}]$. Moreover, 1 is an eigenvalue of infinite multiplicity. To see this take $q = -\operatorname{div} \nabla\phi$ with any ϕ smooth and compactly supported in Ω. Hence $\nabla\phi$ is in $\mathbf{W}_\mathrm{D}^{1,2}$, the domain of $\mathbf{A} = -\operatorname{div} \nabla_{|\mathbf{W}_\mathrm{D}^{1,2}}$, and

$$S^\infty q = \operatorname{div} \mathbf{A}^{-1} \nabla \operatorname{div} \nabla\phi = \operatorname{div} \mathbf{A}^{-1} \operatorname{div} \nabla\nabla\phi = -\operatorname{div} \nabla\phi = q. \tag{2.16}$$

This is the complete description of $\sigma_{\mathrm{ess}}(S^\infty)$ in this case. It is not known whether $\beta(\Omega, \mathrm{D}) = \inf \sigma_{\mathrm{ess}}(S^\infty)$. Surprisingly, the exact value of $\beta(\Omega, \mathrm{D})$ is not known even for the unit square domain. The study also provides

2.2 A priori estimates and invertibility of the PCD operator

By (2.5) we defined $X_{\alpha,\mathbf{w},\Gamma}^{-1} \in \mathcal{L}((W_\Gamma^{1,2})^\#)$. But the first immediate question is whether this operator (continuously) maps L^2 functions to L^2 functions. In the following lemmas we give various possible conditions for ensuring this.

Lemma 2.5 (L^2-bound with $W^{1,3+\epsilon}$ Laplacian regularity)**.** *Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz domain, $\Gamma \subset \partial\Omega$ open, $\alpha \in [0,1]$, and $\mathbf{w} \in \mathbf{W}^{1,2}$. Furthermore, let there exist $\epsilon \geq 0$ when $\alpha = 0$, $\epsilon > 0$ when $\alpha \in (0,1]$ such that A_Γ^{-1} maps L^2 continuously into $W_\Gamma^{1,3+\epsilon}$. Then $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ maps L^2 continuously into L^2.*

Moreover

$$\begin{aligned}\|X_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}(L^2)} \leq 1 &+ \|\mathbf{w}\|_6 \, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3})} \\ &+ \alpha \,\|\operatorname{div}\mathbf{w}\|_2 \, C_{\mathrm{P}}(3+\epsilon,\Omega,\Gamma)\, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3+\epsilon})}.\end{aligned} \tag{2.17}$$

Corollary 2.6 (L^2-bound with Ω creased Lipschitz domain)**.** *Let $\Omega \subset \mathbb{R}^3$ with Dirichlet boundary Γ and Neumann boundary $\partial\Omega\backslash\Gamma$ is a creased Lipschitz domain, $\alpha \in [0,1]$, and $\mathbf{w} \in \mathbf{W}^{1,2}$. Then $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ maps L^2 continuously into L^2.*

Furthermore there exists $C(\Omega,\Gamma) > 0$ such that

$$\|X_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}(L^2)} \leq 1 + C(\Omega,\Gamma)\ (\|\mathbf{w}\|_6 + \alpha\|\operatorname{div}\mathbf{w}\|_2)\,. \tag{2.18}$$

Lemma 2.7 (L^2-bound with $\|\mathbf{w}\|_\infty$ and $\|\operatorname{div}\mathbf{w}\|_3$ bound)**.** *Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz domain, $\Gamma \subset \partial\Omega$ is open, $\alpha \in [0,1]$, and $\mathbf{w} \in L^\infty$ with $\operatorname{div}\mathbf{w} \in L^3$. Then $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ maps L^2 continuously into L^2.*

Moreover

$$\begin{aligned}\|X_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}(L^2)} &\leq 1 + \|\mathbf{w}\|_\infty \, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,2})} \\ &\qquad + \alpha \,\|\operatorname{div}\mathbf{w}\|_3 \, C_{\mathrm{P}}(2,\Omega,\Gamma)\, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,2})} \\ &\leq 1 + |\Omega|^{\frac{1}{3}}\, C_{\mathrm{P}}(2,\Omega,\Gamma)\,(\|\mathbf{w}\|_\infty + C_{\mathrm{P}}(2,\Omega,\Gamma)\,\alpha\,\|\operatorname{div}\mathbf{w}\|_3)\,.\end{aligned} \tag{2.19}$$

Proof of lemmas and corollary. The assertions in the lemmas follow using Hölder's inequality and the Poincaré-Sobolev inequalities. The last inequality in (2.19) holds due to

$$\|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,2})} \leq |\Omega|^{\frac{1}{3}}\, C_{\mathrm{P}}(2,\Omega,\Gamma)\|A_\Gamma^{-1}\|_{\mathcal{L}((W_\Gamma^{1,2})^\#,W_\Gamma^{1,2})} = |\Omega|^{\frac{1}{3}}\, C_{\mathrm{P}}(2,\Omega,\Gamma).$$

The corollary is then a consequence of $W^{1,3+\epsilon}$-theory for creased Lipschitz domains, see (2.2), respectively. □

Employing the $W^{1,3+\epsilon}$-estimate, valid on creased Lipschitz domains, to assure boundedness of $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ in $\mathcal{L}(L^2)$ uniformly in $\|\mathbf{w}\|_{1,2}$, as presented in Corollary 2.6, seems to be new. It is interesting that this choice of function spaces and Lebesgue exponents works just sharply in spatial dimension 3.

In the following we characterize when $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is injective as a mapping in $\mathcal{L}((W_\Gamma^{1,2})^\#)$. Notice that boundedness in $\mathcal{L}(L^2)$ is not needed in particular.

Lemma 2.8 (Injectivity of $X_{\alpha,\mathbf{w},\Gamma}^{-1}$)**.** *Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz domain, $\Gamma \subset \partial\Omega$ be open, $\alpha \in [0,1]$, and $\mathbf{w} \in \mathbf{W}^{1,2}$ with $\mathbf{w}\cdot\mathbf{n} \geq 0$ on $\partial\Omega\backslash\Gamma$. Furthermore assume that*

$$\left|\alpha - \tfrac{1}{2}\right| \, \|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}} < \frac{1}{C_{\mathrm{P}}(2,\Omega,\Gamma)^2}. \tag{2.20}$$

Then $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is injective on $(W_\Gamma^{1,2})^\#$.

some description of $\sigma_{\mathrm{ess}}(S^\infty)$ for three-dimensional corner domains. A description of $\sigma_{\mathrm{ess}}(S^\infty)$ for general D is missing, but it seems plausible that corners belonging to the interior of D will, at least qualitatively, contribute in the same way, as the related eigenfunctions are highly localized. On the other hand, the situation around $\partial\Omega \setminus \mathrm{D}$ might be quite different.

Proof. A_Γ^{-1} is injective on $(W_\Gamma^{1,2})^\#$ and $F_{\alpha,\mathbf{w},\Gamma}$ is injective on $A_\Gamma^{-1}((W_\Gamma^{1,2})^\#) = W_\Gamma^{1,2}$ under the assumptions by Lemma 2.1. □

Now we can establish invertibility of $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ in $\mathcal{L}(L^2)$.

Theorem 2.9. *Assume $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is in $\mathcal{L}(L^2)$ (as for example assured by one of Lemma 2.5, Corollary 2.6, or Lemma 2.7). Furthermore let us assume that the conditions of Lemma 2.8, in particular* (2.20)*, are met. Then $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ maps L^2 onto itself. Hence the inverse operator $X_{\alpha,\mathbf{w},\Gamma}$ exists and is continuous in $\mathcal{L}(L^2)$.*

Proof. $X_{\alpha,\mathbf{w},\Gamma}^{-1} - I = K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1}$ is compact in $\mathcal{L}(L^2)$; see (2.5). Hence from its injectivity the operator $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is surjective by the Fredholm alternative, and hence invertible. Moreover, by the bounded inverse theorem (Theorem A.1), its inverse is also bounded. □

We conclude that Theorem 2.9 ensures that $X_{\alpha,\mathbf{w},\Gamma}$ exists, mapping L^2 continuously onto itself. But it does not give us a bound on the norm $\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}$ in terms of the data, in particular in certain norms of $\mathbf{w}$ and $\operatorname{div}\mathbf{w}$. Our goal will now be to get dependence only on the norm of $\mathbf{w}$ in the natural energy space $\mathbf{W}^{1,2}$ and eventually on the smallness of $\operatorname{div}\mathbf{w}$ but definitely avoiding a dependence on $\|\mathbf{w}\|_\infty$ and $\|\operatorname{div}\mathbf{w}\|_3$ (or even the assumption $\operatorname{div}\mathbf{w} = 0$) which has appeared in most of the literature. But first we start with the case with the restrictive $\|\mathbf{w}\|_\infty + \alpha\|\operatorname{div}\mathbf{w}\|_3$ dependence for comparison with published results. Then we deal with situations in which $\mathbf{w}$ is small in some sense allowing for especially simple treatment. We finish this section by providing bounds which depend solely on $\|\mathbf{w}\|_6$ and $\alpha\|\operatorname{div}\mathbf{w}\|_2$.

Wind controlled in L^∞ and wind divergence in L^3 (restrictive case)

In the preceding section we gave sufficient conditions for the existence of $X_{\alpha,\mathbf{w},\Gamma}$ continuous on L^2. Assume now this is the case. Then we can write

$$X_{\alpha,\mathbf{w},\Gamma} = I - K_{\alpha,\mathbf{w},\Gamma}F_{\alpha,\mathbf{w},\Gamma}^{-1} \tag{2.21}$$

which is simply confirmed by checking (2.5). Assuming further that $A_\Gamma^{-1} \in \mathcal{L}(L^2, W_\Gamma^{1,q})$ (note that $q = 3+\epsilon, 3, 2$ were important cases in Lemma 2.5, 2.7, and Corollary 2.6) and using the formula

$$F_{\alpha,\mathbf{w},\Gamma}^{-1} = A_\Gamma^{-1}X_{\alpha,\mathbf{w},\Gamma}, \tag{2.22}$$

we conclude that $F_{\alpha,\mathbf{w},\Gamma}^{-1}$ maps L^2 into $W_\Gamma^{1,q}$.

It is not obvious that the bounds hold in the case $q = 3+\epsilon, 3$. On the other hand for $q = 2$, the Laplacian solve A_Γ^{-1} is bounded by the standard theory and the convection-diffusion solve $F_{\alpha,\mathbf{w},\Gamma}^{-1}$ is bounded by the ellipticity estimate of Lemma 2.1. This is counter-balanced by the need for a more severe bound on $K_{\alpha,\mathbf{w},\Gamma}$ resulting in strong requirements on the wind $\mathbf{w}$. Now we formulate this precisely in the following theorem.

Theorem 2.10. *Suppose the assumptions of Theorem 2.9 are satisfied. Then the following bound holds:*

$$\begin{aligned} \|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)} \le 1 &+ C_\mathrm{P}(2,\Omega,\Gamma)\,|\Omega|^{\frac{1}{3}}\,\left(\|\mathbf{w}\|_\infty + \alpha\, C_\mathrm{P}(2,\Omega,\Gamma)\,\|\operatorname{div}\mathbf{w}\|_3\right) \\ &\times \left(1 - \left|\alpha - \tfrac{1}{2}\right|\, C_\mathrm{P}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\right)^{-1}. \end{aligned} \tag{2.23}$$

Proof. By the assumptions, $X_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(L^2)$ and formula (2.21) holds. We can express the norm $\|F_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}((W_\Gamma^{1,2})^\#,W_\Gamma^{1,2})}$ as

$$\begin{aligned}\|F_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}((W_\Gamma^{1,2})^\#,W_\Gamma^{1,2})}^{-1} &= \left[\sup_{f\in(W_\Gamma^{1,2})^\#}\frac{\|\nabla F_{\alpha,\mathbf{w},\Gamma}^{-1}f\|_2}{\|f\|_{(W_\Gamma^{1,2})^\#}}\right]^{-1}\\ &= \inf_{f\in(W_\Gamma^{1,2})^\#}\frac{\|f\|_{(W_\Gamma^{1,2})^\#}}{\|\nabla F_{\alpha,\mathbf{w},\Gamma}^{-1}f\|_2} = \inf_{r\in W_\Gamma^{1,2}}\frac{\|F_{\alpha,\mathbf{w},\Gamma}r\|_{(W_\Gamma^{1,2})^\#}}{\|\nabla r\|_2}\\ &= \inf_{r\in W_\Gamma^{1,2}}\sup_{s\in W_\Gamma^{1,2}}\frac{\langle F_{\alpha,\mathbf{w},\Gamma}r,s\rangle}{\|\nabla r\|_2\|\nabla s\|_2} \geq \inf_{r\in W_\Gamma^{1,2}}\frac{\langle F_{\alpha,\mathbf{w},\Gamma}r,r\rangle}{\|\nabla r\|_2^2}\\ &\geq 1-\left|\alpha-\tfrac{1}{2}\right| C_{\mathrm{P}}(2,\Omega,\Gamma)^2\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}},\end{aligned} \tag{2.24}$$

where the third equality holds due to the surjectivity of $F_{\alpha,\mathbf{w},\Gamma}\in\mathcal{L}(W_\Gamma^{1,2},(W_\Gamma^{1,2})^\#)$ by the standard Lax-Milgram theory. The last inequality follows from Lemma 2.1. Combining the embedding

$$\|r\|_2\leq|\Omega|^{\frac{1}{3}}\|r\|_6\leq C_{\mathrm{P}}(2,\Omega,\Gamma)|\Omega|^{\frac{1}{3}}\|\nabla r\|_2 \qquad \text{for all } r\in W_\Gamma^{1,2}$$

with (2.24), we get

$$\|F_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,2})}\leq C_{\mathrm{P}}(2,\Omega,\Gamma)\,|\Omega|^{\frac{1}{3}}\left(1-\left|\alpha-\tfrac{1}{2}\right|C_{\mathrm{P}}(2,\Omega,\Gamma)^2\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\right)^{-1}. \tag{2.25}$$

For the convective term we have the estimate

$$\|K_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(W_\Gamma^{1,2},L^2)}\leq\|\mathbf{w}\|_\infty+\alpha C_{\mathrm{P}}(2,\Omega,\Gamma)\,\|\operatorname{div}\mathbf{w}\|_3. \tag{2.26}$$

Noticing the formula (2.21) and using estimates (2.25) and (2.26) we obtain the desired estimate. □

Small wind

Another circumstance in which the bounds on $\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}$ can be derived is the situation of small data. By the formula (2.5) it is obvious that for $\mathbf{w}$ small enough, $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is invertible, $X_{\alpha,\mathbf{w},\Gamma}$ can be expressed by a Neumann series, and norm $\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}$ is bounded by a geometric series (of numbers). This will be formulated precisely in the following lemma.

Lemma 2.11 (Small data). *Assume $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is in $\mathcal{L}(L^2)$ (as for example assured by one of Lemma 2.5, Corollary 2.6, or Lemma 2.7).*

If

$$\|K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1}\|_{\mathcal{L}(L^2)}<1, \tag{2.27}$$

then $X_{\alpha,\mathbf{w},\Gamma}\in\mathcal{L}(L^2)$ exists, is given by formula

$$X_{\alpha,\mathbf{w},\Gamma}=\sum_{k=0}^{\infty}\left(-K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1}\right)^k, \tag{2.28}$$

and

$$\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}\leq\left(1-\|K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1}\|_{\mathcal{L}(L^2)}\right)^{-1}. \tag{2.29}$$

If in particular, any of the bounds (2.17), (2.18), or (2.19) hold with a right-hand side smaller than 2, then (2.27) is fulfilled, and if we denote the respective right-hand side by rhs it holds

$$\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}\leq(1-(rhs-1))^{-1}. \tag{2.30}$$

Now we can express the smallness condition (2.27) in terms of $\|\mathbf{w}\|_6$ and $\|\operatorname{div}\mathbf{w}\|_2$ norms as in Lemma 2.5.

Corollary 2.12. *Assume the conditions of Lemma 2.5 are met. Further assume that*

$$\|\mathbf{w}\|_6 \, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3})} + \alpha \, \|\operatorname{div}\mathbf{w}\|_2 \, C_\mathrm{P}(3+\epsilon,\Omega,\Gamma) \, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3+\epsilon})} < 1. \tag{2.31}$$

Then $X_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(L^2)$ exists, is given by formula (2.28), *and*

$$\begin{aligned}\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)} \le \Big(1 &- \|\mathbf{w}\|_6 \, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3})} \\ &- \alpha \, \|\operatorname{div}\mathbf{w}\|_2 \, C_\mathrm{P}(3+\epsilon,\Omega,\Gamma) \, \|A_\Gamma^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3+\epsilon})} \Big)^{-1}.\end{aligned} \tag{2.32}$$

In particular, under the conditions of Corollary 2.6, if

$$\|\mathbf{w}\|_6 + \alpha \|\operatorname{div}\mathbf{w}\|_2 < C(\Omega,\Gamma)^{-1} \tag{2.33}$$

with $C(\Omega,\Gamma)$ from Corollary 2.6, then $X_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(L^2)$ exists, is given by formula (2.28), *and*

$$\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)} \le \Big(1 - C(\Omega,\Gamma) \, (\|\mathbf{w}\|_6 + \alpha \|\operatorname{div}\mathbf{w}\|_2) \Big)^{-1}. \tag{2.34}$$

Now we briefly mention a stronger mode of smallness which allows one to obtain field-of-values bounds for the composition $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$.

Observation 2.13 (Very small data)**.** *Let $\mathbf{F} \in \mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$ satisfy estimate* (2.6)*. Assume $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ is in $\mathcal{L}(L^2)$ (as for example assured by one of Lemma 2.5, Corollary 2.6, or Lemma 2.7).*

If

$$\|K_{\alpha,\mathbf{w},\Gamma} A_\Gamma^{-1}\|_{\mathcal{L}(L^2)} < \frac{\underline{S}}{\|S\|_{\mathcal{L}(L^2)}}, \tag{2.35}$$

then $X_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(L^2)$ exists and is given by formula (2.28)*, estimate* (2.29) *holds, and*

$$\inf_{x \in L^2} \frac{\Big\langle SX_{\alpha,\mathbf{w},\Gamma}^{-1} \, x, x \Big\rangle}{\|x\|_2^2} \ge \underline{S} - \|S\|_{\mathcal{L}(L^2)} \|K_{\alpha,\mathbf{w},\Gamma} A_\Gamma^{-1}\|_{\mathcal{L}(L^2)} > 0. \tag{2.36}$$

Proof. This easily follows from Lemma 2.3. □

Wind controlled in L^6 and its divergence in L^2 (the case with no restriction)

Now we focus on the case with less restrictive conditions on wind $\mathbf{w}$. We want to avoid a dependence on norms $\|\mathbf{w}\|_\infty$ and $\|\operatorname{div}\mathbf{w}\|_3$ appearing in (2.23). Under the conditions of Lemma 2.5 (or Corollary 2.6) and Theorem 2.9 we know by formula (2.22) that $F_{\alpha,\mathbf{w},\Gamma}^{-1} \in \mathcal{L}(L^2, W_\Gamma^{1,3+\epsilon})$ and $X_{\alpha,\mathbf{w},\Gamma} \in \mathcal{L}(L^2)$ but we do not know yet how the operator norm $\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}$ depends on $\mathbf{w}$.

Lemma 2.14 ($W^{1,3}$-estimate for convection-diffusion solution)**.** *Let $\alpha \in [0,1]$ and $\epsilon \in [0, \frac{3}{\sqrt{5}})$ be fixed. Let Ω be a bounded Lipschitz domain such that $A_\Gamma^{-1} \in \mathcal{L}((W_\Gamma^{1,\frac{3+\epsilon}{2+\epsilon}})^\#, W_\Gamma^{1,3+\epsilon})$. Let $\mathbf{w} \in \mathbf{W}^{1,2}$ be such that $F_{\alpha,\mathbf{w},\Gamma}^{-1} \in \mathcal{L}(L^2, W_\Gamma^{1,2})$. Denote*

$$\begin{aligned} C_{A_\Gamma^{-1}}(3+\epsilon) &:= \|A_\Gamma^{-1}\|_{\mathcal{L}((W_\Gamma^{1,\frac{3+\epsilon}{2+\epsilon}})^\#, W_\Gamma^{1,3+\epsilon})} \\ C_{F_{\alpha,\mathbf{w},\Gamma}^{-1}} &:= \|F_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,2})}. \end{aligned}$$

It holds that

$$\begin{aligned}\|F_{\alpha,\mathbf{w},\Gamma}^{-1}\|_{\mathcal{L}(L^2,W_\Gamma^{1,3})} \le {} & C_{A_\Gamma^{-1}}(3) \, C_\mathrm{P}(\tfrac{3}{2},\Omega,\Gamma) \\ & \times \Big[|\Omega|^{\frac{1}{6}} + (\|\mathbf{w}\|_6 + C_\mathrm{P}(2,\Omega,\Gamma) \, \alpha \, \|\operatorname{div}\mathbf{w}\|_2) \, C_{F_{\alpha,\mathbf{w},\Gamma}^{-1}} \Big].\end{aligned} \tag{2.37}$$

If in addition $\epsilon \in (0, \frac{3}{\sqrt{5}})$*, then*

$$
\begin{aligned}
&\|F^{-1}_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2,W^{1,3+\epsilon}_\Gamma)} \leq \\
&\quad \tfrac{(1+\epsilon)(1+\frac{\epsilon}{3})}{(1+\frac{\sqrt{5}}{3})(1-\frac{\sqrt{5}}{3})}\, C_{A_\Gamma^{-1}}(3+\epsilon)\, C_{\mathrm{P}}(\tfrac{3+\epsilon}{2+\epsilon},\Omega,\Gamma)\, |\Omega|^{\frac{1}{6}\frac{3-\epsilon}{3+\epsilon}} \\
&\quad + \tfrac{(1+\frac{\epsilon}{3})^2}{(1+\frac{\sqrt{5}}{3})(1-\frac{\sqrt{5}}{3})} \left[C_{A_\Gamma^{-1}}(3+\epsilon)\, C_{\mathrm{P}}(\tfrac{3+\epsilon}{2+\epsilon},\Omega,\Gamma)\, \|\mathbf{w}\|_6\right]^{\frac{1+\epsilon}{1+\frac{\epsilon}{3}}} \; C_{F^{-1}_{\alpha,\mathbf{w},\Gamma}} \\
&\quad + \tfrac{(1+\epsilon)(1-\frac{\epsilon}{3})}{(1+\frac{\sqrt{5}}{3})(1-\frac{\sqrt{5}}{3})}\, C_{\mathrm{P}}(2,\Omega,\Gamma)\, C_{\mathrm{P}}(3+\epsilon,\Omega,\Gamma)^{\frac{\frac{2}{3}\epsilon}{1-\frac{\epsilon}{3}}} \\
&\quad \times \left[C_{A_\Gamma^{-1}}(3+\epsilon)\, C_{\mathrm{P}}(\tfrac{3+\epsilon}{2+\epsilon},\Omega,\Gamma)\, \alpha\, \|\operatorname{div}\mathbf{w}\|_2\right]^{\frac{1+\frac{\epsilon}{3}}{1-\frac{\epsilon}{3}}} \; C_{F^{-1}_{\alpha,\mathbf{w},\Gamma}}.
\end{aligned}
\tag{2.38}
$$

Proof. Fix $f \in L^2$. By the assumptions there exists $u \in W^{1,2}$ such that $f = F_{\alpha,\mathbf{w},\Gamma} u = A_\Gamma u + K_{\alpha,\mathbf{w},\Gamma} u$ and $\|\nabla u\|_2 \leq C_{F^{-1}_{\alpha,\mathbf{w},\Gamma}} \|f\|_2$. We can therefore write $u = A_\Gamma^{-1}(f - K_{\alpha,\mathbf{w},\Gamma} u) = A_\Gamma^{-1} f - A_\Gamma^{-1}(\mathbf{w}\cdot\nabla u) - A_\Gamma^{-1}(u\,\alpha\operatorname{div}\mathbf{w})$. The first term can be estimated by

$$\|\nabla A_\Gamma^{-1} f\|_{3+\epsilon} \leq |\Omega|^{\frac{1}{6}\frac{3-\epsilon}{3+\epsilon}}\, C_{\mathrm{P}}(\tfrac{3+\epsilon}{2+\epsilon},\Omega,\Gamma)\, C_{A_\Gamma^{-1}}(3+\epsilon). \tag{2.39}$$

The second term can be estimated by

$$\|\nabla A_\Gamma^{-1}(\mathbf{w}\cdot\nabla u)\|_{3+\epsilon} \leq C_{A_\Gamma^{-1}}(3+\epsilon)\, \|\mathbf{w}\|_6\, C_{\mathrm{P}}(\tfrac{3+\epsilon}{2+\epsilon},\Omega,\Gamma)\, \|\nabla u\|_2^{1-\mu}\, \|\nabla u\|_{3+\epsilon}^{\mu}, \tag{2.40}$$

where we used the interpolation $\|\nabla u\|_{(\frac{1-\mu}{2}+\frac{\mu}{3+\epsilon})^{-1}} \leq \|\nabla u\|_2^{1-\mu}\, \|\nabla u\|_{3+\epsilon}^{\mu}$ and Hölder's inequality assuming $\frac{1-\mu}{2} + \frac{\mu}{3+\epsilon} + \frac{1}{6} + \frac{1}{(3+\epsilon)'^*} = 1$. This requirement gives $\mu = \frac{2}{3}\frac{\epsilon}{1+\epsilon}$ and shows that the estimate works whenever $\epsilon \geq 0$. The third term can be estimated by

$$
\begin{aligned}
\|\nabla A_\Gamma^{-1}(u\,\alpha\operatorname{div}\mathbf{w})\|_{3+\epsilon} &\leq C_{A_\Gamma^{-1}}(3+\epsilon)\, \alpha\, \|\operatorname{div}\mathbf{w}\|_2\, C_{\mathrm{P}}(\tfrac{3+\epsilon}{2+\epsilon},\Omega,\Gamma) \\
&\quad \times (C_{\mathrm{P}}(2,\Omega,\Gamma)\|\nabla u\|_2)^{1-\lambda}\, (C_{\mathrm{P}}(3+\epsilon,\Omega,\Gamma)\|\nabla u\|_{3+\epsilon})^{\lambda}
\end{aligned}
\tag{2.41}
$$

with the interpolation $\|u\|_{\frac{6}{1-\lambda}} \leq \|u\|_6^{1-\lambda}\, \|u\|_\infty^{\lambda}$ and Hölder's inequality assuming $\frac{1-\lambda}{6} + \frac{1}{2} + \frac{1}{(3+\epsilon)'^*} = 1$. This condition gives $\lambda = \frac{2}{3}\frac{\epsilon}{1+\frac{\epsilon}{3}}$, which is in the range $(0,1)$ for $\epsilon \in (0,3)$ and is $\lambda = 0$ when $\epsilon = 0$. Hence the term $\|u\|_\infty$ does not appear in the case $\epsilon = 0$, when the embedding to L^∞ does not hold. Collecting (2.39), (2.40), and (2.41) gives

$$\|\nabla u\|_{3+\epsilon} \leq C_0 + C_1 \|\nabla u\|_2^{1-\mu}\, \|\nabla u\|_{3+\epsilon}^{\mu} + C_2 \|\nabla u\|_2^{1-\lambda}\, \|\nabla u\|_{3+\epsilon}^{\lambda} \tag{2.42}$$

with C_0, C_1, and C_2 given by the quantities from (2.39), (2.40), and (2.41). Using Young's inequality and subtracting gives

$$(1-\mu-\lambda)\, \|\nabla u\|_{3+\epsilon} \leq C_0 + \left[(1-\mu)\, C_1^{\frac{1}{1-\mu}} + (1-\lambda)\, C_2^{\frac{1}{1-\lambda}}\right] \|\nabla u\|_2, \tag{2.43}$$

which is the desired estimate if $1-\mu-\lambda > 0$ which happens if $\epsilon < \frac{3}{\sqrt{5}}$. This proves (2.38). The estimate (2.37) is a special case for $\epsilon = 0$. □

Theorem 2.15. *Let* Ω *be a bounded Lipschitz domain,* $\Gamma \subset \partial\Omega$ *be an open domain, and* $\alpha \in [0,1]$*. Assume that* $A_\Gamma^{-1} \in \mathcal{L}((W_\Gamma^{1,\frac{3+\epsilon}{2+\epsilon}})^\#, W_\Gamma^{1,3+\epsilon})$ *for some* $\epsilon \in (0, \frac{3}{\sqrt{5}})$ *if* $\alpha = 0$ *or for some* $\epsilon \in [0, \frac{3}{\sqrt{5}})$ *if* $\alpha \in (0,1]$*. Let* $\mathbf{w} \in \mathbf{W}^{1,2}$ *be such that* $\mathbf{w}\cdot\mathbf{n} \geq 0$ *on* $\partial\Omega\backslash\Gamma$*. Also assume that* $\operatorname{div}\mathbf{w}$ *is sufficiently small in the sense of* (2.20)*.*

Then there exist positive constants $C_0(\Omega,\Gamma)$*,* $C_1(\Omega,\Gamma)$*,* $C_2(\Omega,\Gamma)$*,* $C_3(\Omega,\Gamma)$*,* $C_4(\Omega,\Gamma)$*,*

$C_5(\Omega,\Gamma)$, $1 < \Lambda_0(\Omega,\Gamma) < 1 + \frac{\sqrt{5}-1}{2}$, $1 < \Lambda_1(\Omega,\Gamma) < 2 + \frac{\sqrt{5}-1}{2}$ *such that*

$$\begin{aligned} \|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)} \le 1 &+ C_0(\Omega,\Gamma)\,\|\mathbf{w}\|_6 + C_1(\Omega,\Gamma)\,\alpha\,\|\operatorname{div}\mathbf{w}\|_2 \\ &+ \Big[C_2(\Omega,\Gamma)\,\|\mathbf{w}\|_6^2 + C_3(\Omega,\Gamma)\,\alpha\,\|\operatorname{div}\mathbf{w}\|_2\,\|\mathbf{w}\|_6 \\ &\quad + C_4(\Omega,\Gamma)\,\alpha\,\|\operatorname{div}\mathbf{w}\|_2\,\|\mathbf{w}\|_6^{\Lambda_0(\Omega,\Gamma)} \\ &\quad + C_5(\Omega,\Gamma)\,(\alpha\,\|\operatorname{div}\mathbf{w}\|_2)^{1+\Lambda_1(\Omega,\Gamma)}\Big] \\ &\quad \times \Big(1 - \big|\alpha - \tfrac{1}{2}\big|\; C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\Big)^{-1}. \end{aligned} \tag{2.44}$$

Proof. Using the formula (2.21) we can estimate

$$\begin{aligned} \|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)} \le 1 &+ \|\mathbf{w}\|_6\,\|F^{-1}_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2,W^{1,3}_\Gamma)} \\ &+ \alpha\,\|\operatorname{div}\mathbf{w}\|_2\,C_{\mathrm{P}}(3+\epsilon,\Omega,\Gamma)\,\|F^{-1}_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2,W^{1,3+\epsilon}_\Gamma)}. \end{aligned} \tag{2.45}$$

By Lemma 2.1, the Sobolev-Poincaré embedding, and Hölder's inequality, it holds that

$$\|F^{-1}_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2,W^{1,2}_\Gamma)} \le |\Omega|^{\frac{1}{3}}\,C_{\mathrm{P}}(2,\Omega,\Gamma)\,\Big(1 - \big|\alpha-\tfrac{1}{2}\big|\;C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\Big)^{-1}. \tag{2.46}$$

Combining estimates from Lemma 2.14 with (2.45) and (2.46) gives the desired estimate. Constants $\Lambda_0(\Omega,\Gamma) := \frac{1+\epsilon}{1+\frac{\epsilon}{3}}$ and $\Lambda_1(\Omega,\Gamma) := \frac{1+\frac{\epsilon}{3}}{1-\frac{\epsilon}{3}}$ are bounded from above by $1+\frac{\sqrt{5}-1}{2}$ and $2+\frac{\sqrt{5}-1}{2}$, respectively, due to the requirement $\epsilon < \frac{3}{\sqrt{5}}$.

When $\alpha = 0$, estimates (2.45) and (2.37) show that $\epsilon = 0$ is sufficient. □

Summary of obtained a priori bounds

Above we have obtained bounds on $\|X^{-1}_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}$ and $\|X_{\alpha,\mathbf{w},\Gamma}\|_{\mathcal{L}(L^2)}$ under different possible circumstances. Using Theorem A.5, these bounds immediately imply bounds on the spectrum of $X^{-1}_{\alpha,\mathbf{w},\Gamma}$, and in combination with Lemma 2.3, also imply bounds on the spectrum of $SX^{-1}_{\alpha,\mathbf{w},\Gamma}$. More precisely, using appropriate statements from this section, we have

$$\sigma(SX^{-1}_{\alpha,\mathbf{w},\Gamma}) \subset \{\lambda\in\mathbb{C} : \tfrac{1}{C_1} \le |\lambda| \le C_2\} \tag{2.47}$$

with certain $C_1,\ C_2 > 0$ depending continuously on

$$\begin{aligned} C_1 &= C_1\Big(\Omega,\mathrm{D},\Gamma,\alpha,\mathbf{w},\big(1-|\alpha-\tfrac{1}{2}|\,C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\big)^{-1},\underline{\mathbf{F}}^{-1}\|\mathbf{F}\|^2_{\mathcal{L}(\mathbf{W}^{1,2}_{\mathrm{D}},(\mathbf{W}^{1,2}_{\mathrm{D}})^{\#})}\Big), \\ C_2 &= C_2\big(\Omega,\mathrm{D},\Gamma,\alpha,\mathbf{w},\underline{\mathbf{F}}^{-1}\big). \end{aligned} \tag{2.48}$$

The dependence on $(\alpha,\mathbf{w})$ is either through $\|\mathbf{w}\|_\infty + \alpha\|\operatorname{div}\mathbf{w}\|_3$ or $\|\mathbf{w}\|_6 + \alpha\|\operatorname{div}\mathbf{w}\|_2$ if Ω is a creased Lipschitz domain; see the preceding statements in this section. Note that the latter is controlled by $\|\mathbf{w}\|_2 + \|\nabla\mathbf{w}\|_2$, the norm of $\mathbf{w}$ in the natural energy space $\mathbf{W}^{1,2}$. The dependence on $\big(1-|\alpha-\frac{1}{2}|\,C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\big)^{-1}$ expresses the fact that $\frac{1}{C_1} \to 0+$ as $|\alpha-\frac{1}{2}|\,C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}} \to 1-$; see Theorem 2.15 or 2.10. Hence either the smallness of $\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}$ or the skew-symmetric convective term $\alpha = \frac{1}{2}$ is needed.

Although this section did not provide useful quantitative bounds, which would, besides other things, explain qualities of the preconditioner and its superiority to the "Stokes preconditioner" $S^{-1} \approx I$, it is nevertheless a prerequisite for any subsequent analysis to have uniform bounds in reasonable norms of the data. For example, for creased Lipschitz domains we know that the norms of the preconditioned Schur complement (and in turn its spectrum) are bounded uniformly in data size $\|\mathbf{w}\|_{\mathbf{W}^{1,2}}$.

Notice that we have not yet consider any mesh discretization and we have worked in the respective function spaces. Hence one might hope (and prove) that the uniform bounds eventually transfer to discretized operators, for example by finite element methods, and will stay uniform with mesh refinements under certain conditions, for example, quasi-uniform refinement or even adaptive schemes.[2] We will provide some uniform bounds in the discrete case in Theorem 3.3, Remark 3.4, and Remark 3.5.

We note that Deuring [15] obtained spectral bounds for a discrete version of a certain modification of $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ with dependence on $\|\mathbf{w}\|_\infty + \|\operatorname{div}\mathbf{w}\|_3$. Most other published spectral estimates build on the result of Loghin [43] which requires $\operatorname{div}\mathbf{w} = 0$ and features bounds which depend on $\|\mathbf{w}\|_\infty$. We will discuss these results in more detail in Section 3.5, p. 120.

2.3 GMRES iterations with the preconditioned saddle-point system

Let us consider the operator Q representing a linearized Navier-Stokes problem, for example the Oseen system (1.2), given by

$$Q \coloneqq \begin{pmatrix} \mathbf{F} & \nabla \\ -\operatorname{div} & 0 \end{pmatrix}. \tag{2.49}$$

Remember that we assumed compactness of $\mathbf{K} = \mathbf{F} - \mathbf{A}$ in $\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$. Now consider a right preconditioner which approximates the ideal preconditioner (1.7) and is given by

$$\hat{P}_\mp \coloneqq \begin{pmatrix} \mathbf{F} & \nabla \\ 0 & \mp X_{\alpha,\mathbf{w},\Gamma} \end{pmatrix}, \tag{2.50}$$

where we can choose one of the signs $\mp$. The two variants of the preconditioned operator read

$$Q\hat{P}_\mp^{-1} = \begin{pmatrix} I & 0 \\ -\operatorname{div}\mathbf{F}^{-1} & \pm SX_{\alpha,\mathbf{w},\Gamma}^{-1} \end{pmatrix}. \tag{2.51}$$

Due to the preceding exposition we already know that under certain conditions this operator is well defined, specifically it is bounded in $\mathcal{L}((\mathbf{W}_\mathrm{D}^{1,2})^\# \times L^2)$. In this section we will show how convergence of the GMRES method (described in Appendix B) applied to the saddle-point operator $Q\hat{P}_\mp^{-1}$ can be bounded by the behavior of GMRES iterations with $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$.

As was already pointed out, for operator (1.9) a solution is found by GMRES after at most two iterations, i.e., $r_k = 0$ holds for all $k \geq 2$, because for any polynomial p divisible by $(1-t)(1\mp t)$ it holds that $p(QP_\mp^{-1}) = 0$. For the approximation $Q\hat{P}_\mp^{-1}$ given by (2.51) this obviously does not hold. Assume a polynomial p of a finite degree with coefficients given by $p(t) = \sum_k a_k t^k$. It is straightforward to check that

$$p(Q\hat{P}_\mp^{-1}) = \begin{pmatrix} p(1)I & 0 \\ \sum_k a_k \sum_{j=0}^{k-1} T^j(-\operatorname{div}\mathbf{F}^{-1}) & p(T) \end{pmatrix}, \tag{2.52}$$

where $T \coloneqq \pm SX_{\alpha,\mathbf{w},\Gamma}^{-1}$. Now formally, we would like to use $\sum_{j=0}^{k-1} T^j = (I - T^k)(I-T)^{-1}$, where I is the identity operator on L^2, but this is not possible because it is not guaranteed that $I - T$ is invertible. If it were, then $\sum_k a_k \sum_{j=0}^{k-1} T^j = (p(1)I - p(T))(I-T)^{-1}$, and it could be conluded that, for any initial residual $r_0 = \begin{pmatrix} r_0^\mathbf{v} \\ r_0^p \end{pmatrix} \in (\mathbf{W}_\mathrm{D}^{1,2})^\# \times L^2$,

$$\begin{aligned} &\min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \frac{\left\|p(Q\hat{P}_\mp^{-1})r_0\right\|_{(\mathbf{W}_\mathrm{D}^{1,2})^\#\times L^2}}{\|r_0\|_{(\mathbf{W}_\mathrm{D}^{1,2})^\#\times L^2}} \leq \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1 \\ p(1)=0}} \frac{\left\|p(T)\big((I-T)^{-1}\operatorname{div}\mathbf{F}^{-1}r_0^\mathbf{v} + r_0^p\big)\right\|_2}{\|r_0\|_{(\mathbf{W}_\mathrm{D}^{1,2})^\#\times L^2}} \\ &\leq C(Q,\hat{P}_\mp) \sup_{z\in L^2} \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1 \\ p(1)=0}} \frac{\|p(T)z\|_2}{\|z\|_2} \leq C(Q,\hat{P}_\mp)\|I-T\|_{\mathcal{L}(L^2)} \sup_{z\in L^2} \min_{\substack{p\in\mathcal{P}_{k-1} \\ p(0)=1}} \frac{\|p(T)z\|_2}{\|z\|_2} \end{aligned} \tag{2.53}$$

with constant $C(Q,\hat{P}_\mp)$ depending on the operators but independent of k and r_0. This shows that the behavior of GMRES with operator $Q\hat{P}_\mp^{-1}$ and any initial residual r_0 is controlled by the behavior of the worst-case GMRES[5] with opearator T with a lag of one iteration. Unfortunately, as $I - T$ is not guaranteed to be invertible, $C(Q,\hat{P}_\mp)$ can be infinite. We will provide a remedy to this problem after we look into the spectral properties of $T = \pm SX_{\alpha,\mathbf{w},\Gamma}^{-1}$.

[5] See [40, 61, 24] for details about worst-case GMRES.

Now we look into the structure of the operator $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$. We can write the Schur complement S as a compact perturbation of the Stokes Schur complement $S^\infty := -\operatorname{div}\mathbf{A}^{-1}\nabla$. Let us recall the assumption that the velocity convective term $\mathbf{K} = \mathbf{F} - \mathbf{A}$ is compact in $\mathcal{L}(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$. Then indeed,

$$\begin{aligned} S - S^\infty &= -\operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\nabla + \operatorname{div}\mathbf{A}^{-1}\nabla \\ &= -\operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}(\mathbf{A}+\mathbf{K}-\mathbf{K})\mathbf{A}^{-1}\nabla + \operatorname{div}\mathbf{A}^{-1}\nabla \\ &= \operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\mathbf{K}\mathbf{A}^{-1}\nabla \end{aligned}$$

which is compact in $\mathcal{L}(L^2)$. As a consequence we get

$$\begin{aligned} SX_{\alpha,\mathbf{w},\Gamma}^{-1} &= S(I + K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1}) = S^\infty + (S - S^\infty) + SK_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1} \\ &= S^\infty + \operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\mathbf{K}\mathbf{A}^{-1}\nabla - \operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\nabla K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1}. \end{aligned}$$

Hence

$$SX_{\alpha,\mathbf{w},\Gamma}^{-1} = S^\infty - \operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\mathcal{C} \tag{2.54a}$$

$$\mathcal{C} = \nabla K_{\alpha,\mathbf{w},\Gamma}A_\Gamma^{-1} - \mathbf{K}\mathbf{A}^{-1}\nabla \tag{2.54b}$$

with $\mathcal{C}$ compact in $\mathcal{L}(L^2, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$. As a consequence, using Theorem A.10, we get that $\sigma_\mathrm{ess}(SX_{\alpha,\mathbf{w},\Gamma}^{-1}) = \sigma_\mathrm{ess}(S^\infty)$. By (2.15) and Theorem A.7 we have

$$\overline{\operatorname{Num}}(S^\infty) = [\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]. \tag{2.55}$$

Thus, by virtue of Theorem A.10, the spectrum of $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ consists of $\sigma_\mathrm{ess}(S^\infty) \subset \overline{\operatorname{Num}}(S^\infty) = [\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]$ and at most countably many isolated eigenvalues with finite geometric and algebraic multiplicities. In particular, for $T = -SX_{\alpha,\mathbf{w},\Gamma}^{-1}$, either 1 is not in the spectrum of T or it is an isolated eigenvalue of finite geometric multiplicity. In any case, thanks to Theorem A.8 we can define, with $\gamma_1 \subset \mathbb{C}$ a sufficiently tight Jordan loop around $\{1\}$, the spectral projection

$$P_1 := -\frac{1}{2\pi i}\int_{\gamma_1}(T - zI)^{-1}\,\mathrm{d}z, \tag{2.56}$$

which is a bounded projection on L^2 commuting with T; for T_1 defined as a restriction of T to P_1L^2 and T_2 a restriction of T to $(I-P_1)L^2$ it holds that $\sigma(T_1) = \{1\}$ (if P_1 is non-trivial) and $\sigma(T_2) = \sigma(T)\setminus\{1\}$. As a consequence P_1L^2 and $(I-P_1)L^2$ are invariant subspaces of T and they directly sum to L^2. Hence $T = T_1P_1 + T_2(I-P_1)$ and

$$\begin{aligned} \sum_{j=0}^{k-1} T^j &= \sum_{j=0}^{k-1} T_1^jP_1 + \sum_{j=0}^{k-1} T_2^j(I-P_1) \\ &= \sum_{j=0}^{k-1} T_1^jP_1 + (I_2 - T_2^k)(I_2 - T_2)^{-1}(I-P_1) \end{aligned}$$

where I_2 is the identity operator on $(I-P_1)L^2$. We can continue with computing the term in (2.52):

$$\begin{aligned} \sum_k a_k \sum_{j=0}^{k-1} T^j &= \sum_k a_k \sum_{j=0}^{k-1} T_1^jP_1 + \Big(\sum_k a_kI_2 - \sum_k a_kT_2^k\Big)(I_2 - T_2)^{-1}(I-P_1) \\ &= \Big(\sum_k a_k \sum_{j=0}^{k} T_1^j - \sum_k a_kT_1^k\Big)P_1 \\ &\quad + \Big(\sum_k a_kI_2 - \sum_k a_kT_2^k\Big)(I_2 - T_2)^{-1}(I-P_1) \\ &= \sum_k a_k \sum_{j=0}^{k} T_1^jP_1 + \sum_k a_kI_2(I_2 - T_2)^{-1}(I-P_1) && (2.57) \\ &\quad - p(T)(P_1 + (I_2 - T_2)^{-1}(I-P_1)). && (2.58) \end{aligned}$$

Now we can express T_1 in Jordan canonical form as $T_1 = I_1 + N$ where N is a direct sum of left shifts and I_1 is the identity operator on $P_1 L^2$. Using the binomial theorem we have, on $P_1 L^2$,

$$T_1^j = \sum_{i=0}^{j} \binom{j}{i} N^i,$$

$$\sum_{j=0}^{k} T_1^j = \sum_{j=0}^{k} \sum_{i=0}^{j} \binom{j}{i} N^i = \sum_{i=0}^{k} N^i \sum_{j=i}^{k} \binom{j}{i} = \sum_{i=0}^{k} \binom{k+1}{i+1} N^i,$$

where we used the hockey-stick identity[6] in the last equality. Denoting the length of the longest Jordan chain in T_1 by $L < \infty$, we have $N^L = 0$ and

$$\sum_k a_k \sum_{j=0}^{k} T_1^j = \sum_k a_k \sum_{i=0}^{k} \binom{k+1}{i+1} N^i = \sum_{i=0}^{L-1} N^i \sum_{k=i}^{\infty} \binom{k+1}{i+1} a_k. \tag{2.59}$$

A requirement that all terms in (2.57) vanish can be equivalently expressed using (2.59) as

$$\sum_k \binom{k+1}{i} a_k = 0 \qquad \text{for all } i = 0, 1, \dots, L \tag{2.60}$$

with conventions that $\binom{k+1}{i} = 0$ for $k + 1 < i$ and $L = 0$ when $P_1 = 0$. Under (2.60) the terms in (2.57) vanish and we have

$$\sum_k a_k \sum_{j=0}^{k-1} T^j = -p(T)(P_1 + (I_2 - T_2)^{-1}(I - P_1)). \tag{2.61}$$

So now we can estimate the k-th GMRES residual using (2.52) and (2.61)

$$\begin{aligned} &\min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \frac{\left\|p(Q\hat{P}_+^{-1})r_0\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}}{\left\|r_0\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}} \\ &\le \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1\\ (2.60)}} \frac{\left\|p(T)\big((P_1 + (I_2 - T_2)^{-1}(I - P_1))\operatorname{div}\mathbf{F}^{-1}r_0^{\mathbf{v}} + r_0^p\big)\right\|_2}{\left\|r_0\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}} \\ &\qquad\qquad \le C(Q, \hat{P}_+) \sup_{z\in L^2} \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1\\ (2.60)}} \frac{\|p(T)z\|_2}{\|z\|_2}. \end{aligned} \tag{2.62}$$

It is not clear to us whether the argument used to derive (2.62) works for $Q\hat{P}_-^{-1}$; $T = SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ has 1 in its essential spectrum and it is an eigenvalue of infinite multiplicity. If it is an isolated point of the spectrum, which is true at least for a Lipschitz $\Omega \subset \mathbb{R}^2$ in the no-slip situation $\mathrm{D} = \partial\Omega$, see [12, Theorem 3.3], then the spectral projector (2.56) can be defined; the estimate (2.62) is then valid provided $L < \infty$, which we cannot verify.

Now we would like to estimate the minimum on the right-hand side of (2.62) by

$$\min_{\substack{p\in\mathcal{P}_{k-(L+1)}\\ p(0)=1}} \frac{\|p(T)z\|_2}{\|z\|_2}$$

as in (2.53). Consider certain polynomials with coefficients given by summands of (2.60); specifically consider that for $i = 1, 2, \dots$

$$\begin{aligned} \sum_k \tbinom{k+1}{i} a_k t^{k-i+1} &= \sum_k \left(\tbinom{k}{i} + \tbinom{k}{i-1}\right) a_k t^{k-i+1} \\ = \sum_k a_k \tbinom{k}{i} t^{k-i} t + \sum_k a_k \tbinom{k}{i-1} t^{k-(i-1)} &= \tfrac{p^{(i)}(t)}{i!} t + \tfrac{p^{(i-1)}(t)}{(i-1)!} \end{aligned} \tag{2.63a}$$

[6] or the Christmas stocking identity, depending on reader's geocultural preference

and for $i = 0$

$$\sum_k \tbinom{k+1}{i} a_k t^{k-i+1} = p(t)t. \tag{2.63b}$$

Hence (2.60) is fulfilled whenever $p^{(i)}(1) = 0$ for all $i = 0, 1, \dots, L$. This is clearly true when $p(t) = (1-t)^{L+1} q(t)$ with any polynomial q. Moreover $q(0) = 1$ implies that $p(0) = 1$ and we can estimate

$$\min_{\substack{p\in\mathcal{P}_k\\ p(0)=1\\ (2.60)}} \|p(T)z\|_2 \le \|I - T\|^{L+1}_{\mathcal{L}(L^2)} \min_{\substack{p\in\mathcal{P}_{k-(L+1)}\\ p(0)=1}} \|p(T)z\|_2.$$

We will put these estimates together in the following theorem.

Theorem 2.16. *Assume that the conditions of Lemma 2.3, Lemma 2.4, and Theorem 2.9 are met.*

Operator $-SX^{-1}_{\alpha,\mathbf{w},\Gamma}$ either does not have a neighborhood of 1 in its spectrum or 1 is an isolated eigenvalue of finite algebraic multiplicity. Denote by L the length of the longest Jordan chain in $-SX^{-1}_{\alpha,\mathbf{w},\Gamma}$ associated with the eigenvalue 1, i.e., L is the smallest non-negative integer such that $(I + SX^{-1}_{\alpha,\mathbf{w},\Gamma})^L P_1 = 0$, where P_1 is the spectral projector associated with operator $-SX^{-1}_{\alpha,\mathbf{w},\Gamma}$ and eigenvalue 1.

Then there exists $C(Q, \hat{P}_+) > 0$ such that the residuals generated by GMRES with operator $Q\hat{P}_+^{-1}$ and any initial residual $r_0 \in (\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2$ and the worst-case GMRES with operator $SX^{-1}_{\alpha,\mathbf{w},\Gamma}$ are related by

$$\min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \frac{\left\|p(Q\hat{P}_+^{-1})r_0\right\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\#\times L^2}}{\|r_0\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\#\times L^2}} \le C(Q,\hat{P}_+) \sup_{z\in L^2} \min_{\substack{p\in\mathcal{P}_{k-(L+1)}\\ p(0)=1}} \frac{\left\|p(SX^{-1}_{\alpha,\mathbf{w},\Gamma})z\right\|_2}{\|z\|_2} \tag{2.64}$$

for all $k \ge L+1$. The constant $C(Q, \hat{P}_+)$ is finite and can be bounded by

$$\begin{aligned} C(Q,\hat{P}_+) &= \sup_{\substack{(z^{\mathbf{v}},z^p)\in(\mathbf{W}^{1,2}_{\mathrm{D}})^\#\times L^2\\ \|(z^{\mathbf{v}},z^p)\|=1}} \left\|\left(I + SX^{-1}_{\alpha,\mathbf{w},\Gamma}\right)^L \left(\operatorname{div}\mathbf{F}^{-1}z^{\mathbf{v}} + \left(I + SX^{-1}_{\alpha,\mathbf{w},\Gamma}\right)z^p\right)\right\|_2 \\ &\le \left(C_+^\infty + \Sigma(\Omega,\mathrm{D})\underline{\mathbf{F}}^{-1}\|\mathcal{C}\|_{\mathcal{L}(L^2,(\mathbf{W}^{1,2}_{\mathrm{D}})^\#}\right)^L \\ &\quad \times \left(C_+^\infty + \Sigma(\Omega,\mathrm{D})\underline{\mathbf{F}}^{-1}\left(1 + \|\mathcal{C}\|_{\mathcal{L}(L^2,(\mathbf{W}^{1,2}_{\mathrm{D}})^\#)}\right)\right), \end{aligned} \tag{2.65}$$

$$C_+^\infty = \Sigma(\Omega,\mathrm{D})^2 + 1. \tag{2.66}$$

Proof. The first part of the theorem, the existence of $L < \infty$, follows from the theorems in Appendix A as discussed above. Now we show (2.64) with the specific definition of $C(Q, \hat{P}_+)$. Denote $T = -SX^{-1}_{\alpha,\mathbf{w},\Gamma}$; define T_1, T_2 as a restriction of T to P_1L^2, $(I-P_1)L^2$, respectively; denote by I_1, I_2 the identity operator on P_1L^2, $(I-P_1)L^2$, respectively. Consider formula (2.52), which is valid for any polynomial p. For $p(t) = \sum_k a_k t^k$ subject to (2.60) we have, by (2.52) and (2.61),

$$p(Q\hat{P}_+^{-1}) = \begin{pmatrix} 0 & 0 \\ p(T)\bar{P}\operatorname{div}\mathbf{F}^{-1} & p(T) \end{pmatrix} \tag{2.67}$$

with $\bar{P} = P_1 + (I_2 - T_2)^{-1}(I - P_1)$. Now consider, with polynomial q, $p(t) = (1-t)^{L+1}q(t)$. By (2.63) such p automatically fulfills (2.60). With such p, using the direct sums $T = T_1 + T_2$, $I = I_1 + I_2$,

$$\begin{aligned} p(T)\bar{P} &= q(T)(I_1 + I_2 - T_1 - T_2)^{L+1}(P_1 + (I_2 - T_2)^{-1}(I - P_1)) \\ &= q(T)((I_1 - T_1)^{L+1}P_1 + (I_2 - T_2)^L(I - P_1)), \end{aligned}$$

but due to the definition of L we have $(I_1 - T_1)^L = 0$, and we get

$$p(T)\bar{P} = q(T)(I_2 - T_2)^L(I - P_1) = q(T)(I - T)^L. \tag{2.68}$$

Using (2.67) and (2.68) we obtain for any $r_0 = \begin{pmatrix} r_0^{\mathbf{v}} \\ r_0^p \end{pmatrix} \in (\mathbf{W}_{\mathrm{D}}^{1,2})^{\#} \times L^2$

$$\begin{aligned}
\min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \frac{\left\|p(Q\hat{P}_+^{-1})r_0\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}}{\left\|r_0\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}} &\le \min_{\substack{q\in\mathcal{P}_{k-(L+1)}\\ q(0)=1}} \frac{\left\|q(T)(I-T)^L\big(\operatorname{div}\mathbf{F}^{-1}r_0^{\mathbf{v}} + (I-T)r_0^p\big)\right\|_2}{\left\|r_0\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}} \\
&\le \sup_{(z^{\mathbf{v}},z^p)\in(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2} \frac{\left\|(I-T)^L\big(\operatorname{div}\mathbf{F}^{-1}z^{\mathbf{v}} + (I-T)z^p\big)\right\|_2}{\left\|(z^{\mathbf{v}},z^p)\right\|_{(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2}} \\
&\quad\times \sup_{(z^{\mathbf{v}},z^p)\in(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2} \min_{\substack{q\in\mathcal{P}_{k-(L+1)}\\ q(0)=1}} \frac{\left\|q(T)(I-T)^L\big(\operatorname{div}\mathbf{F}^{-1}z^{\mathbf{v}} + (I-T)z^p\big)\right\|_2}{\left\|(I-T)^L\big(\operatorname{div}\mathbf{F}^{-1}z^{\mathbf{v}} + (I-T)z^p\big)\right\|_2} \\
&\le C(Q,\hat{P}_+) \sup_{z\in L^2} \min_{\substack{q\in\mathcal{P}_{k-(L+1)}\\ q(0)=1}} \frac{\|q(T)z\|_2}{\|z\|_2}
\end{aligned}$$

so that (2.64) and the equality in (2.65) are proved. Employing formula (2.54), we have

$$\begin{aligned}
I - T = I + SX_{\alpha,\mathbf{w},\Gamma}^{-1} &= I + S^\infty - \operatorname{div}\mathbf{F}^{-1}\mathcal{C}, \\
\|I-T\|_{\mathcal{L}(L^2)} &\le \|I + S^\infty\|_{\mathcal{L}(L^2)} + \Sigma(\Omega,\mathrm{D})\|\mathbf{F}^{-1}\|_{\mathcal{L}((\mathbf{W}_{\mathrm{D}}^{1,2})^{\#},\mathbf{W}_{\mathrm{D}}^{1,2})}\|\mathcal{C}\|_{\mathcal{L}(L^2,(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#})} \\
&\le 1 + \Sigma(\Omega,\mathrm{D})^2 + \Sigma(\Omega,\mathrm{D})\underline{\mathbf{F}}^{-1}\|\mathcal{C}\|_{\mathcal{L}(L^2,(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#})}
\end{aligned}$$

which follows from (2.12) and (2.55). This proves the inequality in (2.65). □

Remark. *Theorem 2.16 does not apply directly to the other case $\hat{P}_-$, where $T = SX_{\alpha,\mathbf{w},\Gamma}^{-1}$, because it is not known to the author whether $T_1 - I_1$ is in this case nilpotent. Nevertheless this variant, corresponding to the minus sign in* (2.50)*, is often seen to perform better in practice. It is interesting to notice that, when $T_1 - I_1$ is nilpotent, precisely $(SX_{\alpha,\mathbf{w},\Gamma}^{-1} - I)^L P_1 = 0$, then the estimate analogous to* (2.64) *holds for $\hat{P}_-$ with the constant*

$$\begin{aligned}
C(Q,\hat{P}_-) &= \sup_{\substack{(z^{\mathbf{v}},z^p)\in(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}\times L^2\\ \|(z^{\mathbf{v}},z^p)\|=1}} \left\|\big(I - SX_{\alpha,\mathbf{w},\Gamma}^{-1}\big)^L\Big(\operatorname{div}\mathbf{F}^{-1}z^{\mathbf{v}} + \big(I - SX_{\alpha,\mathbf{w},\Gamma}^{-1}\big)z^p\Big)\right\|_2 \\
&\le \Big(C_-^\infty + \Sigma(\Omega,\mathrm{D})\underline{\mathbf{F}}^{-1}\|\mathcal{C}\|_{\mathcal{L}(L^2,(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#}}\Big)^L \\
&\quad\times \Big(C_-^\infty + \Sigma(\Omega,\mathrm{D})\underline{\mathbf{F}}^{-1}\big(1 + \|\mathcal{C}\|_{\mathcal{L}(L^2,(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#})}\big)\Big), \\
C_-^\infty &= \max\{\Sigma(\Omega,\mathrm{D})^2 - 1,\, 1 - \beta(\Omega,\mathrm{D})^2\},
\end{aligned}$$

which is better than (2.65) *due to $C_-^\infty < C_+^\infty$. This is an indication of better peformance of $\hat{P}_-$, at least for small convection, assuming L is small.*

2.4 Spectrum of the preconditioned Schur complement

In Section 2.2 we showed that the spectrum of $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ is bounded away from zero and infinity uniformly in a certain norm of the data. In the preceding section we further showed that the spectrum of $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ consists of $\sigma_{\mathrm{ess}}(SX_{\alpha,\mathbf{w},\Gamma}^{-1}) = \sigma_{\mathrm{ess}}(S^\infty) \subset [\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]$ and isolated eigenvalues of finite geometric multiplicity. Now we employ results from operator perturbation theory to conclude that the isolated eigenvalues accumulate at $[\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]$ at a certain rate.

Using Theorem (A.12), (2.55), and (2.54) we obtain

$$\sum_{\lambda\in\sigma_{\mathrm{p}}(SX_{\alpha,\mathbf{w},\Gamma}^{-1})} \operatorname{dist}\bigl(\lambda, [\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]\bigr)^p \\ \leq \|X_{\alpha,\mathbf{w},\Gamma}^{-1} - S^\infty\|_{\mathcal{S}_p(L^2)}^p = \|\operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\mathcal{C}\|_{\mathcal{S}_p(L^2)}^p \tag{2.69}$$

if the right-hand side is finite for certain $p > 1$. Now we show that $\operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\mathcal{C}$ is in the p-Schatten class, see Definition A.11, for some $p > 6$. The first compact term in (2.54b) features the compact embedding $L^2 \hookrightarrow (W_\Gamma^{1,\frac{3+\epsilon}{2+\epsilon}})^\#$. We will now assume structurally similar compactness in $\mathbf{K}$, namely assume that $\mathbf{K}$ is the Picard or Newton linearization of the non-linear velocity convective term $\mathbf{v} \mapsto (1-\alpha)\mathbf{v}\cdot\nabla\mathbf{v} + \alpha\operatorname{div}(\mathbf{v}\otimes\mathbf{v})$. In such a situation $\mathbf{K}$ features the compact embedding $(L^{\frac{3}{2}})^3 \hookrightarrow (\mathbf{W}_\mathrm{D}^{1,2})^\#$. By [62, Theorem 1.107], the approximation numbers of these embeddings defined by (A.9) decay as

$$a_k\bigl(L^2 \hookrightarrow (W_\Gamma^{1,\frac{3+\epsilon}{2+\epsilon}})^\#\bigr) \sim k^{-\frac{1}{6}\frac{1-\frac{\epsilon}{3}}{1+\frac{\epsilon}{3}}}, \tag{2.70a}$$

$$a_k\bigl((L^{\frac{3}{2}})^3 \hookrightarrow (\mathbf{W}_\mathrm{D}^{1,2})^\#\bigr) \sim k^{-\frac{1}{6}}. \tag{2.70b}$$

Considering definition (A.9) and Definition A.11, representation formula (2.54) and decay rates (2.70) imply that $SX_{\alpha,\mathbf{w},\Gamma}^{-1} - S^\infty \in \mathcal{S}_{6+\hat{\varepsilon}}(L^2)$ with any $\hat{\varepsilon} > 0$ because any $\epsilon > 0$ is sufficient to obtain the bounds (2.17). This verifies that the right-hand side of the accumulation rate (2.69) is finite. In the following theorem we give precise conditions for the validity of this estimate.

Theorem 2.17. *Let the conditions of Corollary 2.6 be fulfilled. Furthermore let us assume that the condition of Lemma 2.3, i.e.,* (2.6)*, holds. Furthermore assume that the linearized velocity convection* $\mathbf{K} = \mathbf{F} - \mathbf{A}$ *is in* $\mathcal{S}_6(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$.

Then $SX_{\alpha,\mathbf{w},\Gamma}^{-1} \in \mathcal{L}(L^2)$ *and*

$$\sigma_{\mathrm{ess}}(SX_{\alpha,\mathbf{w},\Gamma}^{-1}) \subset [\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2].$$

Furthermore, with any $p > 6$ *it holds that*

$$\left(\sum_{\lambda\in\sigma_{\mathrm{p}}(SX_{\alpha,\mathbf{w},\Gamma}^{-1})} \operatorname{dist}\bigl(\lambda, [\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]\bigr)^p\right)^{\frac{1}{p}} \leq \Sigma(\Omega,\mathrm{D})\underline{\mathbf{F}}^{-1}\|\mathcal{C}\|_{\mathcal{S}_p(L^2,(\mathbf{W}_\mathrm{D}^{1,2})^\#)} < +\infty$$

where the eigenvalues are counted according to their algebraic multiplicity.

If, in addition, all conditions of Lemma 2.8 are met, then 0 *is not in* $\sigma(SX_{\alpha,\mathbf{w},\Gamma}^{-1})$.

Note that lower bounds on $|\lambda|$ are eventually provided in Section 2.2, for example, in Theorem 2.15, for the situation which is the least restrictive on data. Also note that the condition $\mathbf{K} \in \mathcal{S}_6(\mathbf{W}_\mathrm{D}^{1,2}, (\mathbf{W}_\mathrm{D}^{1,2})^\#)$ is automatically ensured for common linearizations of the velocity convection as pointed out above; in more general situations, e.g., when using streamline-upwind stabilization, etc., the condition has to be verified.

Proof. By Lemma 2.3 and Corrolary 2.6 we have $S, X_{\alpha,\mathbf{w},\Gamma}^{-1} \in \mathcal{L}(L^2)$. Thanks to (2.12), formula (2.54) is valid and $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ is a compact perturbation of S^∞. Hence for the essential spectra $\sigma_{\mathrm{ess}}(SX_{\alpha,\mathbf{w},\Gamma}^{-1}) = \sigma_{\mathrm{ess}}(S^\infty)$, where the latter is contained in $[\beta(\Omega,\mathrm{D})^2, \Sigma(\Omega,\mathrm{D})^2]$ by Lemma 2.3 and Theorem A.5.

Using formulas (2.54) and (2.55), Theorem A.12 and (2.12) give the first inequality. The first term in (2.54b) is estimated as in (2.17) and by (2.2) and (2.70a) it is in $\mathcal{S}_p$ for a certain $p > 6$. As $\epsilon > 0$ can be arbitrarily small, any $p > 6$ is admissible. The second term in (2.54b) is estimated similarly. This bounds $\mathcal{C}$ in $\mathcal{S}_p$ with any $p > 6$ and the proof is finished. □

It is fair to remark that information about the spectrum alone is not sufficient to provide descriptive information about convergence of the GMRES method as showed by Greenbaum, Pták, and Strakoš [26]. Nevertheless we show in the next section that the inclusion $\mathcal{C} \in \mathcal{S}_p$, $p > 6$, together with the self-adjointness of S^∞ implies certain information about GMRES convergence.

2.5 GMRES iterations with the preconditioned Schur complement

In Section 2.3, Theorem 2.16 we showed that the convergence behavior of the preconditioned saddle-point system is governed up to a certain lag by the behavior of GMRES on the preconditioned Schur complement. In this section we employ a new result on GMRES convergence for compactly-perturbed positive self-adjoint operators which we provide in Theorem B.4. But first we make a few remarks on the presence/absence of superlinear convergence for the Stokes operator.

For operators of the form $T = I + C$ with compact C, the superlinear convergence of Krylov subspace iterations has been shown in the literature, i.e., $\|r_k\|/\|r_{k-1}\| \to 0$ with a suitable norm. Moreover a certain rate is guaranteed when C is in some p-Schatten class; see [48] for GMRES, and also [30] and references therein for results on Krylov subspace methods for self-adjoint problems. On the other hand our problem has the structure $T = SX^{-1}_{\alpha,\mathbf{w},\Gamma} = S^\infty + C$ with $S^\infty \in \mathcal{L}(L^2)$ self-adjoint and positive and a compact $C \in \mathcal{S}_{6+\epsilon}(L^2)$. The Stokes Schur complement on Lipschitz domains has fat components in its spectrum; see [12]. Hence no superlinear convergence can be expected for the special case of Stokes; the intervals in the essential spectrum of the Stokes Schur complement would contribute by the Chebyshev polynomials to an eventual composite bound. Note that there appeared incorrect proofs of superlinear convergence of a stabilized Stokes problem; [30, Example 3.9, p. 1320] uses the incorrect assertion that the operator

$$\begin{pmatrix} \mathbf{A}^{-1}\nabla \operatorname{div} & \\ & \operatorname{div}\mathbf{A}^{-1}\nabla \end{pmatrix} : \mathbf{W}^{1,2}_{\partial\Omega} \times L^2/\mathbb{R} \to \mathbf{W}^{1,2}_{\partial\Omega} \times L^2/\mathbb{R}$$

(with $\mathbf{A} = -\Delta_{|\mathbf{W}^{1,2}_{\partial\Omega}}$ in accordance with the previous notation) is compact in $\mathcal{L}(\mathbf{W}^{1,2}_{\partial\Omega} \times L^2/\mathbb{R})$. We will present a convergence result for $T = SX^{-1}_{\alpha,\mathbf{w},\Gamma} = S^\infty + C$ which applies to the Stokes problem as a special case by setting $C \coloneqq 0$.

We now apply Theorem B.4 to the preconditioned Schur complement together with Theorem 2.16 to conclude that GMRES convergence for the preconditioned system $Q\hat{P}_+^{-1}$ is contractive up to a delay which vanishes superlinearly with a certain rate.

Corollary 2.18. *Assume that the conditions of Theorem 2.16 and Theorem 2.17 are met. Then for any $\varepsilon > 0$ there exists $C_\varepsilon \geq 0$ depending on $SX^{-1}_{\alpha,\mathbf{w},\Gamma}$ such that GMRES iterations with operator $Q\hat{P}_+^{-1} \in \mathcal{L}((\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2)$ and initial residual $r_0 \in (\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2$ produce residuals $r_k \in (\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2$ with the norm*

$$\|r_k\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2} = \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(Q\hat{P}_+^{-1})r_0\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2}$$

which fulfils

$$\left(\frac{1}{C(Q,\hat{P}_+)} \frac{\|r_{k+L+1}\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2}}{\|r_0\|_{(\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times L^2}}\right)^{\frac{1}{k}} \leq \frac{1 - \frac{\beta(\Omega,\mathrm{D})^2}{\Sigma(\Omega,\mathrm{D})^2}}{1 + \frac{\beta(\Omega,\mathrm{D})^2}{\Sigma(\Omega,\mathrm{D})^2}} + C_\varepsilon k^{-\frac{1}{6+\varepsilon}} \tag{2.71}$$

with integer $L \geq 0$ and $C(Q,\hat{P}_+) > 0$ from Theorem 2.16; in particular L and $C(Q,\hat{P}_+)$ are independent of r_0.

Note that (2.71) is not sharp for the Stokes case, where $C_\varepsilon = 0$. In that case one could, if having finer information about the spectrum, construct a composite bound covering separately isolated eigenvalues and the essential spectrum. The latter can be treated in the symmetric case by the Chebyshev polynomials which give a better linear bound. Nevertheless in the non-self-adjoint case it is not possible to employ the real Chebyshev polynomials and the first term in estimate (2.71) is analogous to the min-max polynomial on a disk, cf. [39].

2.6 PCD variants and boundary conditions

Equations (1.10) and (1.11) heuristically motivated the two variants of the preconditioner. Furthermore, one might be also interested in simple preconditioning by $S^{-1} \approx X^{-1}_{\alpha,\mathbf{0},\Gamma} = Y^{-1}_{\alpha,\mathbf{0},\Gamma} = I$;

see [19, p. 1300] and references therein. So far our analysis did not cover the Y-variant and we have not specifically discussed the boundary conditions.

Define the Y-variant of the PCD as an adjoint

$$Y_{\alpha,\mathbf{w},\Gamma}^{-1} := X_{1-\alpha,-\mathbf{w},\Gamma}^{-\#}. \tag{2.72}$$

The motivation for the sign of $-\mathbf{w}$ on the right-hand side will become obvious a bit later; $1-\alpha$ instead of α is only for aesthetic reasons which will also become evident. First consider that all the results of Section 2.2, in particular surjectivity, injectivity, norm and spectral estimates for $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ and $X_{\alpha,\mathbf{w},\Gamma}$, are valid also for $X_{1-\alpha,-\mathbf{w},\Gamma}^{-1}$ mutatis mutandis. In particular we know by Theorem 2.9 that, under certain conditions, $X_{1-\alpha,-\mathbf{w},\Gamma}^{-1}$ maps L^2 onto itself, and in turn so does $Y_{\alpha,\mathbf{w},\Gamma}^{-1} = X_{1-\alpha,-\mathbf{w},\Gamma}^{-\#}$. Thus we can safely compute the adjoint

$$Y_{\alpha,\mathbf{w},\Gamma}^{-1} = (I + K_{1-\alpha,-\mathbf{w},\Gamma} A_\Gamma^{-1})^{\#} = I + A_\Gamma^{-1} K_{\alpha,\mathbf{w},\Gamma}^{\mathrm{R}} \tag{2.73}$$

where the operator $K_{\alpha,\mathbf{w},\Gamma}^{\mathrm{R}} = K_{1-\alpha,-\mathbf{w},\Gamma}^{\#}$ is given, in accordance with (2.3), by

$$\begin{gathered}\left\langle K_{\alpha,\mathbf{w},\Gamma}^{\mathrm{R}} r, q\right\rangle = -\int_\Omega \left(\mathbf{w}\cdot\nabla q\, r + (1-\alpha)\operatorname{div}\mathbf{w}\, r\, q\right) \\ \text{whenever } r,q\in W_\Gamma^{1,2} \text{ or } r\in L^2,\ q\in W_\Gamma^{1,3+\epsilon},\ \begin{cases}\epsilon\ge 0 & \text{if } \alpha=0,\\ \epsilon>0 & \text{if } \alpha\in(0,1].\end{cases}\end{gathered} \tag{2.74}$$

The operator (2.74) features a Robin boundary condition $\frac{\partial r}{\partial \mathbf{n}} - \mathbf{w}\cdot\mathbf{n}\, r = 0$ on $\partial\Omega\setminus\Gamma$ in the convection-diffusion solve of

$$r = Y_{\alpha,\mathbf{w},\Gamma}\phi = (A_\Gamma + K_{\alpha,\mathbf{w},\Gamma}^{\mathrm{R}})^{-1} A_\Gamma \phi.$$

To see this, consider that when r is smooth enough, we can integrate by parts in (2.74) to arrive at

$$\left\langle K_{\alpha,\mathbf{w},\Gamma}^{\mathrm{R}} r, q\right\rangle = \int_\Omega \mathbf{w}\cdot\nabla r\, q + \alpha\operatorname{div}\mathbf{w}\, r\, q - \int_{\partial\Omega\setminus\Gamma}\mathbf{w}\cdot\mathbf{n}\, r\, q. \tag{2.75}$$

But considering that the case $r\in L^2$ in (2.74) is important for the definition of (2.73), we have to retain the definition (2.74) rather than (2.75), which is only valid for smooth enough r.

Note also that all the spectral and GMRES convergence results provided in Sections 2.3–2.5 for $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ are valid also for $SY_{\alpha,\mathbf{w},\Gamma}^{-1}$ with obvious modifications. However, note that the commutator formula (2.54) takes a different but structurally similar form

$$\begin{aligned} SY_{\alpha,\mathbf{w},\Gamma}^{-1} &= S^\infty - \operatorname{div}(\mathbf{A}+\mathbf{K})^{-1}\mathcal{C}^Y, \\ \mathcal{C}^Y &= \nabla A_\Gamma^{-1}(K_{\alpha,\mathbf{w},\Gamma}^{\mathrm{R}}) - \mathbf{K}\mathbf{A}^{-1}\nabla. \end{aligned} \tag{2.76}$$

All results which depend on properties of $\mathcal{C}$ in (2.54), e.g., $(6+\varepsilon)$-Schatten compactness, remain true for $\mathcal{C}^Y$.

All results in Sections 2.2–2.3 are also valid for the Stokes preconditioner $S^{-1}\approx X_{\alpha,\mathbf{0},\Gamma}^{-1} = Y_{\alpha,\mathbf{0},\Gamma}^{-1} = I$ simply by setting $\mathbf{w} := \mathbf{0}$ (keep in mind that the underlying problem still can have non-trivial convection $\mathbf{b}$). Note that the right-hand side of (2.64) in this case takes, up to the constant $C(Q,\hat{P}_\mp)$, the form

$$\min_{\substack{p\in\mathcal{P}_{k-(L+1)}\\ p(0)=1}} \|p(S)\|_{\mathcal{L}(L^2)}. \tag{2.77}$$

Thanks to (2.13), which is valid under the condition (2.6), the Schur complement S has its numerical range contained in a half-plane not containing the origin; together with (2.12), these are bounds on the numerical range in terms of $\underline{\mathbf{F}}^{-1}$ and $\|\mathbf{F}\|_{\mathcal{L}(\mathbf{W}_{\mathrm{D}}^{1,2},(\mathbf{W}_{\mathrm{D}}^{1,2})^{\#})}$. The approximation problem (2.77) is tractable using field-of-value estimates; see [35]. It is an interesting point that (2.77) can be estimated (although the estimate might not be tight) in the case of the preconditioner $S^{-1}\approx X_{\alpha,\mathbf{0},\Gamma}^{-1} = Y_{\alpha,\mathbf{0},\Gamma}^{-1} = I$, while in the PCD case $X_{\alpha,\mathbf{w},\Gamma}^{-1}$ (or $Y_{\alpha,\mathbf{w},\Gamma}^{-1}$), the corresponding quantity (2.64) is not tractable using this technique. The price for this is that, as is empirically known, $S^{-1} \approx X_{\alpha,\mathbf{0},\Gamma}^{-1} = Y_{\alpha,\mathbf{0},\Gamma}^{-1} = I$ turns out to be useless as a preconditioner with increasing data size (Reynolds number) and hence any estimates of (2.77) are pointless. We omit any further details for this case.

Now we summarize particular practical choices of the artificial boundary conditions required naturally by the above analysis:

1. preconditioner (2.50) requires for injectivity by Lemma 2.8 that

$$\Gamma \supset \{\mathbf{w} \cdot \mathbf{n} < 0\}; \tag{2.78}$$

2. preconditioner

$$\hat{P}_{\mp} := \begin{pmatrix} \mathbf{F} & \nabla \\ 0 & \mp X_{1-\alpha,-\mathbf{w},\Gamma}^{\#} \end{pmatrix} \tag{2.79}$$

requires for injectivity by Lemma 2.8 that

$$\Gamma \supset \{\mathbf{w} \cdot \mathbf{n} > 0\}; \tag{2.80}$$

we postpone comparison to previously published results to Section 3.5.

2.7 Best-approximation property of the PCD correction

So far we have not considered the synergistic effect of the two compact terms in (2.54b). All aforementioned assertions used the triangle inequality, the worst case estimate, to treat the terms separately. Note that this includes the derivation of $\|\mathcal{C}\|_{\mathcal{S}_{6+\varepsilon}}$ bounds and resulting superlinearly-contractive convergence behavior as stated in Corollary 2.18. The advantage is that the results are valid for the Stokes preconditioner $X_{\alpha,\mathbf{0},\Gamma}^{-1} = I$, which one obtains by setting $\mathbf{w} := \mathbf{0}$ to get $K_{\alpha,\mathbf{w},\Gamma} = 0$. The question is why a particular choice of $\mathbf{w} := \mathbf{b}$ is the best in the sense that the first term in (2.54b) counterbalances with the second term – the term stemming from the fact, that the underlying problem is the Oseen problem rather than the Stokes problem.

In existing studies, a related question has been considered – is the commutator $S - X_{\alpha,\mathbf{w},\Gamma}$ or $S - Y_{\alpha,\mathbf{w},\Gamma}$ small? As of our knowledge the only known answer is that the commutator is zero when $\nabla\mathbf{w} = \mathbf{0}$ and no boundary is present; see [22, section 9.2, equation (9.13)]. This situation is equivalent to

$$S^{\infty} = I, \tag{2.81a}$$

$$\mathcal{C} = 0 \tag{2.81b}$$

in (2.54) which gives $SX_{\alpha,\mathbf{w},\Gamma}^{-1} = I$. This is derived by the same argument as the one used to derive (2.16), which is an interior argument, with no boundary involved. On the other hand it is known that corners of the domain cause the presence of fat essential spectrum in S^{∞}; for simplicity, with a planar polygonal domain and no-slip boudary conditions $\mathrm{D} = \partial\Omega$, a corner of opening ω contributes to the essential spectrum of S^{∞} by $[\frac{1}{2} - \frac{|\sin\omega|}{2\omega}, \frac{1}{2} + \frac{|\sin\omega|}{2\omega}]$; see [12, Theorem 3.3]. Thus validity of (2.81a) cannot be expected and it is clear that this has nothing to do with the convection; this is a feature present even in the Stokes problem on a corner domain. Hence we argue that rather than considering some smallness of $S - X_{\alpha,\mathbf{w},\Gamma}$ or $S - Y_{\alpha,\mathbf{w},\Gamma}$, one should look into the synergy of the two terms in $\mathcal{C}$ because $\mathcal{C}$ is compact so it does not influence the essential spectrum $\sigma_{\mathrm{ess}}(S) = \sigma_{\mathrm{ess}}(S^{\infty})$ and only creates discrete eigenvalues. Without considering any synergy of the two terms in $\mathcal{C}$ we showed that $\mathcal{C}$ is $(6+\varepsilon)$-Schatten; the ultimate goal would be to improve this rate of compactness and thus explain the effectiveness of the preconditioner. Nevertheless we do not know whether this is true.

In what follows we provide Fourier analysis of $\mathcal{C}$ for a very simplified problem. It does not give a definitive answer, but might give some insight. Consider $\Omega = (0,\pi)^2$ and the no-slip problem on the whole boundary, i.e., $\mathrm{D} = \partial\Omega$, $\mathbf{b} = \mathbf{0}$ on $\partial\Omega$. Note that this is a special case which we did not treat as it requires a modified pressure space $L^2/\mathbb{R}$. The appropriate choice of the Dirichlet boundary Γ in A_Γ^{-1} is now $\Gamma = \emptyset$, but A_Γ has to be defined on the space $W^{1,2}/\mathbb{R}$ as it represents a pure Neumann problem. The eigenfunctions of $A_\Gamma = -\operatorname{div}\nabla_{|W^{1,2}/\mathbb{R}}$,

$$\{u_{jk}\}_{j,k=1}^{\infty}, \quad u_{jk} = \cos(jx)\cos(ky),$$

form an orthogonal basis in $L^2/\mathbb{R}$ and $W^{1,2}/\mathbb{R}$; see [44, Appendix A.4]. Analogously, the eigensystem of $\mathbf{A} = -\operatorname{div}\nabla_{|\mathbf{W}_\mathrm{D}^{1,2}}$,

$$\{v_{jk}\mathbf{e}_I\}_{j,k=1,2,\ldots,\, I=1,2}, \quad v_{jk} = \sin(jx)\sin(ky),$$

forms an orthogonal basis of L^2 and $\mathbf{W}_\mathrm{D}^{1,2}$.

Let us choose $\alpha := 0$ and $\mathbf{w} := \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ so that the first term of (2.54b) applied to u_{jk} is

$$\begin{aligned}
\nabla K_{\alpha,\mathbf{w},\Gamma} A_\Gamma^{-1} u_{jk} &= \nabla\big(\mathbf{b}\cdot\nabla \frac{u_{jk}}{j^2+k^2}\big) \\
&= \frac{1}{j^2+k^2}\nabla\big(-jb_1\sin(jx)\cos(ky) - kb_2\cos(jx)\sin(ky)\big) \\
&= \frac{1}{j^2+k^2}\begin{pmatrix} -b_1 j^2 u_{jk} + b_2 jk v_{jk} + \nabla\mathbf{b}(\ldots) \\ -b_2 k^2 u_{jk} + b_1 jk v_{jk} + \nabla\mathbf{b}(\ldots) \end{pmatrix}
\end{aligned} \tag{2.82}$$

where $\nabla\mathbf{b}(\ldots)$ stands for terms proportional to $\nabla\mathbf{b}$. To solve the problem

$$\mathbf{A}(\tilde{u}_{jk}\mathbf{e}_1) = \frac{\partial u_{jk}}{\partial x}\mathbf{e}_1,$$

we expand $\tilde{u}_{jk} = c_{mn}^{jk} v_{mn}$ (summation implied) and test by v_{pq} to obtain

$$\underbrace{c_{mn}^{jk}\int_\Omega \nabla v_{mn}\cdot\nabla v_{pq}}_{(m^2+n^2)\frac{\pi^2}{4}\delta_{mp}\delta_{nq}} = \int_\Omega \frac{\partial u_{jk}}{\partial x} v_{pq} = -j\underbrace{\int_0^\pi \sin(jx)\sin(px)}_{\frac{\pi}{2}\delta_{jp}}\underbrace{\int_0^\pi \cos(ky)\sin(qy)}_{(1-(-1)^{k-q})\frac{q}{q^2-k^2}},$$

and hence

$$c_{mn}^{jk} = \delta_{mp}\delta_{nq}\delta_{jp}\frac{-2j}{\pi(m^2+n^2)}(1-(-1)^{k-q})\frac{q}{q^2-k^2}.$$

Putting this together we have

$$\mathbf{A}^{-1}\frac{\partial u_{jk}}{\partial x}\mathbf{e}_1 = \tilde{u}_{jk}\mathbf{e}_1 = c_{mn}^{jk} v_{mn}\mathbf{e}_1 = -\frac{2j}{\pi}\sin(jx)\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}}\frac{q}{q^2-k^2}\frac{\sin(qy)}{q^2+j^2}\mathbf{e}_1$$

with an analogous formula for $\mathbf{A}^{-1}\frac{\partial u_{jk}}{\partial y}\mathbf{e}_2$ with an appropriate change of indices. Hence we can express

$$\begin{aligned}
&\mathbf{K}\mathbf{A}^{-1}\nabla u_{jk} := \mathbf{b}\cdot\nabla\mathbf{A}^{-1}\nabla u_{jk} \\
&\quad= \begin{pmatrix} -b_1\frac{2j^2}{\pi}\cos(jx)\displaystyle\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}}\frac{q}{q^2-k^2}\frac{\sin(qy)}{q^2+j^2} - b_2\frac{2j}{\pi}\sin(jx)\displaystyle\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}}\frac{q^2}{q^2-k^2}\frac{\cos(qy)}{q^2+j^2} \\ \text{(c.p.)} \end{pmatrix}
\end{aligned} \tag{2.83}$$

where (c.p.) stands for the appropriate cyclic permutation of the indices and coordinates. On the other hand using the formulas

$$\cos(ky) = \frac{2}{\pi}\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}}\frac{q}{q^2-k^2}\sin(qy), \qquad \sin(ky) = -\frac{2}{\pi}\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}}\frac{k}{q^2-k^2}\cos(qy)$$

we can expand (2.82) as

$$\begin{aligned}
&\nabla K_{\alpha,\mathbf{w},\Gamma} A_\Gamma^{-1} u_{jk} \\
&\quad= \begin{pmatrix} -b_1 \tfrac{2j^2}{\pi} \cos(jx) \displaystyle\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}} \tfrac{q}{q^2-k^2} \tfrac{\sin(qy)}{j^2+k^2} - b_2 \tfrac{2j}{\pi} \sin(jx) \displaystyle\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}} \tfrac{k^2}{q^2-k^2} \tfrac{\cos(qy)}{j^2+k^2} \\ \text{(c.p.)} \end{pmatrix} \\
&\qquad + \nabla\mathbf{b}(\ldots). \qquad (2.84)
\end{aligned}$$

Using (2.84) and (2.83) we obtain

$$\begin{aligned}
&\mathcal{C} u_{jk} = \nabla K_{\alpha,\mathbf{w},\Gamma} A_\Gamma^{-1} u_{jk} - \mathbf{K}\mathbf{A}^{-1} \nabla u_{jk} \\
&\quad= \begin{pmatrix} -b_1 \tfrac{2j^2}{\pi(j^2+k^2)} \cos(jx) \displaystyle\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}} \tfrac{q\sin(qy)}{q^2+j^2} - b_2 \tfrac{2j^2}{\pi(j^2+k^2)} \sin(jx) \displaystyle\sum_{\substack{q\in\mathbb{Z}\\ q-k \text{ odd}}} \tfrac{j\cos(qy)}{q^2+j^2} \\ \text{(c.p.)} \end{pmatrix} \\
&\qquad + \nabla\mathbf{b}(\ldots). \qquad (2.85)
\end{aligned}$$

Using the formulas

$$\begin{aligned}
-\frac{\sinh(j(y-\frac{\pi}{2}))}{\sinh(j\frac{\pi}{2})} &= \tfrac{2}{\pi} \sum_{\substack{q\in\mathbb{Z}\\ q \text{ even}}} \frac{q\sin(qy)}{q^2+j^2}, & \frac{\cosh(j(y-\frac{\pi}{2}))}{\sinh(j\frac{\pi}{2})} &= \tfrac{2}{\pi} \sum_{\substack{q\in\mathbb{Z}\\ q \text{ even}}} \frac{j\cos(qy)}{q^2+j^2}, \\
\frac{\cosh(j(y-\frac{\pi}{2}))}{\cosh(j\frac{\pi}{2})} &= \tfrac{2}{\pi} \sum_{\substack{q\in\mathbb{Z}\\ q \text{ odd}}} \frac{q\sin(qy)}{q^2+j^2}, & -\frac{\sinh(j(y-\frac{\pi}{2}))}{\cosh(j\frac{\pi}{2})} &= \tfrac{2}{\pi} \sum_{\substack{q\in\mathbb{Z}\\ q \text{ odd}}} \frac{j\cos(qy)}{q^2+j^2},
\end{aligned}$$

we can simplify (2.85) to

$$\begin{aligned}
&\mathcal{C} u_{jk} - \nabla\mathbf{b}(\ldots) \\
&\quad= \begin{cases} \begin{pmatrix} \tfrac{j^2}{j^2+k^2} \big(\ \ b_1 \cos(jx) \tfrac{\sinh(j(y-\frac{\pi}{2}))}{\sinh(j\frac{\pi}{2})} + b_2 \sin(jx) \tfrac{\cosh(j(y-\frac{\pi}{2}))}{\sinh(j\frac{\pi}{2})} \big) \\ \text{(c.p.)} \end{pmatrix} & k \text{ odd}, \\ \begin{pmatrix} \tfrac{j^2}{j^2+k^2} \big(-b_1 \cos(jx) \tfrac{\cosh(j(y-\frac{\pi}{2}))}{\cosh(j\frac{\pi}{2})} - b_2 \sin(jx) \tfrac{\sinh(j(y-\frac{\pi}{2}))}{\cosh(j\frac{\pi}{2})} \big) \\ \text{(c.p.)} \end{pmatrix} & k \text{ even}. \end{cases} \qquad (2.86)
\end{aligned}$$

It is not clear how to interpret this formula. Nevertheless, by comparing (2.83) and (2.85) we can see that the role of the PCD correction (2.84) is to replace factors $\frac{1}{k^2-q^2}$ and $\frac{q^2}{k^2-q^2}$ by $\frac{1}{j^2+k^2}$ and $\frac{j^2}{j^2+k^2}$, respectively. This possibly achieves the elimination of modes with high j, i.e., high frequencies in the y-direction, which thus ensures the recovery of high frequencies in lower with fewer GMRES iterations. Correspondingly, the hyperbolic factors in (2.86) tend to zero:

$$\frac{\operatorname{cosinh}(j(y-\frac{\pi}{2}))}{\operatorname{cosinh}(j\frac{\pi}{2})} \to 0 \qquad \text{as } j\to\infty \text{ for all } y\in(0,\pi)$$

where cosinh stands for any of sinh, cosh. Nevertheless, the limit is not uniform:

$$\frac{\operatorname{cosinh}(j(y-\frac{\pi}{2}))}{\operatorname{cosinh}(j\frac{\pi}{2})} \to 1 \qquad \text{as } j\to\infty \text{ for } y\in\{0,\pi\}.$$

3 Discretization of the preconditioner

In this section we will propose a discrete realization of the PCD operator in a general setting. This will allow us to introduce the PCD approximation for any L^2-conforming pressure space. In particular, the case of a pressure space using discontinuous elements previously required a special treatment because of the $W^{1,2}$-conformity needed for the PCD Laplacian solve and discretization of the convection operator; see [22, pp. 368–370], which essentially proposes a finite difference scheme for the pressure Laplacian and the pressure convection operator.

In Section 3.1 we propose a general framework which allows the use of a wide range of standard discretization schemes regardless of differentiability-smoothness of the pressure space. Then we apply the framework to three different situations: a $W^{1,2}$-conforming pressure space (Section 3.2), a general L^2-conforming pressure space (Section 3.3), and the piece-wise constant pressure space (Section 3.4), which requires special treatment. As a byproduct we also recover in Section 3.2 the approach for a $W^{1,2}$-conforming pressure space described by Elman, Silvester, and Wathen [21, section 8.2.1] and we identify it with the approximation of the L^2-inner product on the pressure space by its diagonal. This approximation can be avoided for the price of one extra mass matrix solve. In Section 3.5 we provide a comparison of our results with published accounts on PCD.

Note that the framework, and in fact the whole chapter, is largely motivated by the desire to shed some light on correct incorporation of boundary conditions into PCD, an issue which has not been completely understood yet.

> These boundary conditions are not well understood, and a poor choice can critically affect performance. (Elman and Tuminaro [23, p. 257])

We believe that our study could move the problem closer towards correct understanding. We summarize and compare our work to historical results in Section 3.5.

We stress and we are aware that the approaches we propose in this section need numerical testing. A high-perfomance implementation of PCD by Blechta and Řehoř [6] based on the FEniCS project [41, 1], PETSc [2, 3], and petsc4py [13], which has been supported by preliminary results of this work and has been used to study problems in high-performance computing [56, 55], will be used to implement the PCD variants described below in the future.

We note that this study omits completely any issues related to floating point round-off error. Furthermore, efficient and scalable approximations of the mass matrix, the Laplacian, and the convection-diffusion solves are beyond the scope of this work. We refer the interested reader to the monograph by Elman, Silvester, and Wathen [22, section 9.3.3].

From the implementation standpoint it is important to note that it is very common computational practice to use ℓ^2-inner products for GMRES. This is mostly due to the simplicity of existing implementations and the goal of achieving maximal performance. Nevertheless, any performance analysis for the right-preconditioned case uses the problem-dictated inner product on $(\mathbf{W}^{1,2}_{\mathrm{D}})^\# \times (L^2)^\#$, which requires a Laplacian and mass matrix solve in each application. It is seems to be hardly justifiable to replace it by the ℓ^2-inner product, especially on adaptively-refined meshes, although this has not been identified as a possible issue in the literature due to limited test suites mostly working with uniformly and quasi-uniformly refined meshes. We think that this is worth further research and remark that the inner product on $(\mathbf{W}^{1,2}_{\mathrm{D}})^\#$ can be mesh-independently localized due to results of Ciarlet and Vohralík [9] and Blechta, Málek, and Vohralík [5], similarly to the classical approximation of the inner product on $(L^2)^\#$ by its diagonal due to Wathen [63], and thus obtaining a cheap and accurate approximation of the full problem-dictated inner product.

3.1 General considerations

Consider a finite-dimensional pressure space Q^h and a finite-dimensional auxiliary space W^h such that

$$W^h \subset W^{1,\infty}_\Gamma, \qquad\qquad Q^h \subset L^2, \tag{3.1}$$

with bases $\{\phi_j\}_{j=1}^{N^Q}$ and $\{\psi_j\}_{j=1}^{N^W}$, respectively, so that

$$Q^h = \operatorname{span}\{\phi_j\}_{j=1}^{N^Q}, \qquad\qquad W^h = \operatorname{span}\{\psi_j\}_{j=1}^{N^W},$$

and with the inner products

$$(q_1, q_2)_{Q^h} = \langle M^Q q_1, q_2 \rangle_{(Q^h)^\# \times Q^h} \qquad q_1, q_2 \in Q^h, \tag{3.2a}$$

$$(w_1, w_2)_{W^h} = \langle M^W w_1, w_2 \rangle_{(W^h)^\# \times W^h} \qquad w_1, w_2 \in W^h, \tag{3.2b}$$

which we leave so far unspecified. By the Riesz representation theorem the operators $M^Q : Q^h \to (Q^h)^\#$, $M^W : W^h \to (W^h)^\#$ are isometries (in the norms induced by the respective inner products). Furthermore assume that there is a linear transfer operator $\Pi : Q^h \to W^h$. Now define $\Pi^\dagger : W^h \to Q^h$ as the Hilbert adjoint of Π with respect to the inner products (3.2), i.e.,

$$(\Pi^\dagger w, q)_{Q^h} = (\Pi q, w)_{W^h} \qquad \text{for all } q \in Q^h,\ w \in W^h. \tag{3.3}$$

Definition (3.3) can be rewritten using (3.2) as

$$\Pi^\dagger = (M^Q)^{-1}\Pi^\# M^W, \qquad\qquad \Pi = (M^W)^{-1}\Pi^{\dagger\#} M^Q. \tag{3.4}$$

In the following lemma we will characterize the invertibility of $\Pi\Pi^\dagger : W^h \to W^h$.

Lemma 3.1. *The following statements are equivalent:*

(i) $\Pi\Pi^\dagger$ *is invertible,*

(ii) Π *is surjective,*

(iii) $\Pi^\dagger$ *is injective.*

A necessary condition for the validity of (i), (ii), and (iii) is

$$\dim W^h \le \dim Q^h. \tag{3.5}$$

Proof. Implications (i)⇒(ii) and (i)⇒(iii) are obvious.

Choose arbitrary $z \in W^h$ such that $\Pi\Pi^\dagger z = 0$. Then, by setting $w := z$, $q := \Pi^\dagger z$ in (3.3), we obtain

$$0 = (\Pi\Pi^\dagger z, z)_{W^h} = (\Pi^\dagger z, \Pi^\dagger z)_{Q^h} = \|\Pi^\dagger z\|^2_{Q^h},$$

and hence

$$\Pi^\dagger z = 0. \tag{3.6}$$

If (iii) holds, (3.6) implies $z = 0$, so that the kernel of $\Pi\Pi^\dagger$ consists of $\{0\}$ and implication (iii)⇒(i) is proved. On the other hand, (3.3) and (3.6) imply that

$$0 = (\Pi^\dagger z, q)_{Q^h} = (\Pi q, z)_{W^h} \qquad \text{for all } q \in Q^h. \tag{3.7}$$

Assuming (ii) holds, we can choose q in (3.7) such that $\Pi q = z$, so that $0 = (z, z)_{W^h}$. This implies that the kernel of $\Pi\Pi^\dagger$ is $\{0\}$ and implication (ii)⇒(i) is proved.

For proof of the second part of the lemma, consider that by the rank-nullity theorem

$$0 \le \dim\{y \in Q^h,\, \Pi y = 0\} = \dim Q^h - \dim \Pi Q^h.$$

Condition (ii) implies that $\Pi Q^h = W^h$ so that $\dim \Pi Q^h = \dim W^h$ and (3.5) follows. □

Now consider operators $M : Q^h \to (Q^h)^\#$, $A, K, K^{\mathrm{R}} : W^h \to (W^h)^\#$ given, for parameters $\mathbf{w} \in \mathbf{W}^{1,2}$ and $\alpha \in [0, 1]$, by

$$\langle Mp, q\rangle = \int_\Omega p\, q \qquad\qquad p, q \in Q^h, \tag{3.8a}$$

$$\langle Au, v\rangle = \int_\Omega \nabla u \cdot \nabla v \qquad\qquad u, v \in W^h, \tag{3.8b}$$

$$\langle Ku, v\rangle = \int_\Omega \mathbf{w} \cdot \nabla u\, v + \alpha \operatorname{div} \mathbf{w}\, u\, v \qquad\qquad u, v \in W^h, \tag{3.8c}$$

$$\langle K^{\mathrm{R}}u, v\rangle = \int_\Omega \mathbf{w} \cdot \nabla u\, v + \alpha \operatorname{div} \mathbf{w}\, u\, v - \int_{\partial\Omega\setminus\Gamma} \mathbf{w} \cdot \mathbf{n}\, u\, v \qquad\qquad u, v \in W^h. \tag{3.8d}$$

All the integrals are indeed finite thanks to the requirements (3.1). Assuming that $|\Gamma| > 0$ and Π is surjective, we can define the action of the two variants of the discrete PCD operator by

$$X^{-1} := M^{-1}(I + \Pi^{\#} K A^{-1} (\Pi\Pi^{\dagger})^{-\#} \Pi^{\dagger\#}), \tag{3.9a}$$

$$Y^{-1} := (I + \Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} A^{-1} K^{\mathrm{R}} \Pi) M^{-1}. \tag{3.9b}$$

The following theorem shows that the definition is valid and that the operators are under certain conditions invertible.

Theorem 3.2 (Invertibility of discrete PCD operator)**.** *Let $\Omega \subset \mathbb{R}^3$ be a Lipschitz domain and $\Gamma \subset \partial\Omega$ be open with $|\Gamma| > 0$. Let $\mathbf{w} \in \mathbf{W}^{1,2}$, $\alpha \in [0, 1]$. Let spaces* (3.1) *be given. Assume that there is a surjective linear operator $\Pi : Q^h \to W^h$ and let $\Pi^{\dagger} : W^h \to Q^h$ be given by* (3.3)*. Then the definition of $X^{-1}, Y^{-1} : (Q^h)^{\#} \to Q^h$ through* (3.8)*,* (3.9) *is valid.*

Further assume that (2.20) *holds. If $\mathbf{w} \cdot \mathbf{n} \geq 0$ on $\partial\Omega \setminus \Gamma$ then X^{-1} is invertible and its inverse is given by*

$$X = \big(I - \Pi^{\#} K (A + K)^{-1} (\Pi\Pi^{\dagger})^{-\#} \Pi^{\dagger\#}\big) M. \tag{3.10a}$$

If $\mathbf{w} \cdot \mathbf{n} \leq 0$ on $\partial\Omega \setminus \Gamma$ then Y^{-1} is invertible and its inverse is given by

$$Y = M\big(I - \Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} (A + K^{\mathrm{R}})^{-1} K^{\mathrm{R}} \Pi\big). \tag{3.10b}$$

Proof. M^{-1} exists by the Riesz representation theorem. By the surjectivity of Π and Lemma 3.1, the operator $\Pi\Pi^{\dagger}$ is invertible. Thanks to the Dirichlet boundary condition $W^h \subset W^{1,\infty}_{\Gamma}$ with $|\Gamma| > 0$, the Laplace operator A is also invertible by the Riesz representation theorem. Therefore the PCD operators (3.9) are well-defined.

Denote $P := \Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} \Pi$. Obviously $P : Q^h \to Q^h$ is a projector. Note that

$$I = (I - P) + P = (I - P) + \Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} A^{-1} A \Pi.$$

Plugging this into (3.9b) we get

$$Y^{-1} M = (I - P) + \Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} A^{-1} (A + K^{\mathrm{R}}) \Pi. \tag{3.11}$$

A projector on a finite-dimensional space is always continuous, and hence $Q^h = (I - P)Q^h \oplus PQ^h$. The term $(I - P)$ in (3.11) is injective on $(I - P)Q^h$. It remains to show that the second term in (3.11) is injective on PQ^h, which would in turn yield that $Y^{-1}M : Q^h \to Q^h$ is injective and the proof would be finished. First consider that Π is injective on PQ^h. To see this, assume that $\Pi P q = 0$ for some $q \in Q^h$. Hence $0 = \Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} \Pi P q = P^2 q = Pq$ which shows the injectivity of $\Pi_{|PQ^h}$. Next we show that $(A + K^{\mathrm{R}}) : W^h \to (W^h)^{\#}$ is injective. Fix $u \in W^h$ and estimate using integration by parts (noticing that $u = 0$ on Γ and $\mathbf{w} \cdot \mathbf{n} \leq 0$ on $\partial\Omega \setminus \Gamma$), Hölder's inequality, and the Sobolev-Poincaré inequality $\|u\|_6 \leq C_{\mathrm{P}}(2, \Omega, \Gamma) \|\nabla u\|_2$:

$$\begin{aligned}
\langle (A + K^{\mathrm{R}}) u, u \rangle &= \|\nabla u\|_2^2 + \int_{\Omega} \mathbf{w} \cdot \nabla \frac{u^2}{2} + 2\alpha \int_{\Omega} \operatorname{div} \mathbf{w} \, \frac{u^2}{2} - 2 \int_{\partial\Omega\setminus\Gamma} \mathbf{w} \cdot \mathbf{n} \, \frac{u^2}{2} \\
&= \|\nabla u\|_2^2 + (2\alpha - 1) \int_{\Omega} \operatorname{div} \mathbf{w} \, \frac{u^2}{2} - \int_{\partial\Omega\setminus\Gamma} \mathbf{w} \cdot \mathbf{n} \, \frac{u^2}{2} \\
&\geq \|\nabla u\|_2^2 - \frac{|2\alpha - 1|}{2} \|\operatorname{div} \mathbf{w}\|_{\frac{3}{2}} \|u\|_6^2 \\
&\geq \big(1 - |\alpha - \tfrac{1}{2}| C_{\mathrm{P}}(2, \Omega, \Gamma)^2 \|\operatorname{div} \mathbf{w}\|_{\frac{3}{2}}\big) \|\nabla u\|_2^2.
\end{aligned}$$

The parenthetical term in the last line is positive thanks to (2.20). Therefore $A + K^{\mathrm{R}}$ is elliptic, and hence injective. Factors A^{-1} and $(\Pi\Pi^{\dagger})^{-1}$ in (3.11) are injective by construction and $\Pi^{\dagger}$ is injective by the surjectivity of Π and Lemma 3.1. Altogether $\Pi^{\dagger} (\Pi\Pi^{\dagger})^{-1} A^{-1} (A + K^{\mathrm{R}}) \Pi$ is injective on PQ^h and Y^{-1} is hence invertible. The formula (3.10b) is verified against (3.9b) by direct computation of $Y^{-1}Y$; by virtue of the ellipticity of $A + K^{\mathrm{R}}$, the term $(A + K^{\mathrm{R}})^{-1}$ and the formula (3.10b) are well-defined. Hence the proof for the Y^{-1} case is finished. The X^{-1} case is proved in the same way. □

The first reason for the employment of the auxiliary space W^h and the transfer operator $\Pi : W^h \to Q^h$ is that formulas (3.9) are given meaning also when the pressure space Q^h is not regular enough to define the differential operators in (3.8) directly on Q^h, for example when Q^h is a discontinuous finite element space. An alternative to this would be to use a non-conforming discretization for the differential operators in (3.8), possibly including a discontinuous Galerkin method; cf. the finite difference construction in [22, pp. 368–370]. We do not consider this possibility further in this study.

Another reason for the presence of the space W^h is that it incorporates the zero boundary condition on Γ for the Laplacian solve A^{-1} and for the convection-diffusion operator $A + K$ or $A + K^{\mathrm{R}}$ in order to obtain its ellipticity; see the proof of Theorem 3.2.

The significance of Theorem 3.2 is that it guarantees the very first requirement put on the preconditioner – its invertibility. Indeed, a solution p of the underlying problem can be arbitrary in the space Q^h or L^2, in the discrete or continuous case respectively, and hence it must be in the range of the preconditioner for consistency.

The leading-order term M^{-1} in (3.9) is known to be an appropriate preconditioner for the Stokes case $\mathbf{w} = 0$; see [8, 51]. This motivates the construction (3.9) which ensures that the leading term is recovered when $\mathbf{w} = 0$ and that the boundary conditions from W^h do not pollute the leading term.

Now we look at the significance of the wind direction assumption in Theorem 3.2. For a typical non-linear iteration scheme (either Picard or Newton) it is natural to assume that the wind has a correct direction on the inflow boundary, whereas it is difficult to guarantee the sign of the wind on the outflow boundary; cf. the discussion in Section 1. Typically, one would have (1.3) but not (1.4). Hence, using the Y^{-1}-variant with $\Gamma = \Gamma_{\mathrm{out}}$ when (1.3) holds guarantees that the assumption of Theorem 3.2 regarding the wind direction is met. On the other hand, for the X^{-1}-variant with the choice $\Gamma = \Gamma_{\mathrm{in}}$ the wind-direction assumption of the Theorem would be met if (1.4) was satisfied which is difficult to guarantee. This might indicate that the Y-variant is more robust, which might incidentally correspond to its preference in the recent literature [23], [22, Remark 9.3]. Nevertheless our argument indicating robustness of the Y-variant seems to be new. We would like to note that more general cases, e.g., a velocity Dirichlet condition (1.2c) with $\mathbf{v}^{\mathrm{D}} \cdot \mathbf{n} > 0$ on (parts of) Γ_{in}, should be treated slightly differently. Indeed, the subscript $_{\mathrm{in}}$ is not very descriptive in this case as Γ_{in} is no longer exclusively an inflow boundary. It is not difficult to realize that such a case can be treated using the Y-variant with $\Gamma = \Gamma_{\mathrm{out}} \cup \{\mathbf{x} \in \Gamma_{\mathrm{in}},\, \mathbf{v}^{\mathrm{D}}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) > 0\}$ in order to satisfy the wind-direction condition of the theorem.

We will proceed by establishing $\mathcal{L}(L^2)$-bounds on discrete PCD operators X^{-1}, Y^{-1}, and their inverses X, Y, analogously to the infinite-dimensional case in Section 2.2. We start with the case of wind controlled in $\|\mathbf{w}\|_\infty + \alpha \|\operatorname{div} \mathbf{w}\|_3$ and later comment on a difficulty concerning the case $\|\mathbf{w}\|_6 + \alpha \|\operatorname{div} \mathbf{w}\|_2$, which we leave as future work.

Theorem 3.3 (A priori bounds on discrete PCD)**.** *Let the conditions of the first part of Theorem 3.2 be fulfilled. Furthermore, assume that the operators $\Pi : Q^h \to W^h$ and $(\Pi\Pi^\dagger)^{-1}\Pi^\dagger : W^h \to Q^h$ are bounded in $\mathcal{L}(L^2)$, the wind $\mathbf{w} \in L^\infty$, and, if $\alpha > 0$, $\operatorname{div} \mathbf{w} \in L^3$. Then the following bounds hold true:*

$$\begin{aligned} \|X^{-1}\|_{\mathcal{L}\left((L^2)^\#, L^2\right)} &\le 1 + \|\Pi\|_{\mathcal{L}(L^2)} \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)} C_{\mathrm{P}}(2,\Omega,\Gamma)\, |\Omega|^{\frac{1}{3}} \\ &\qquad \times \Big(\|\mathbf{w}\|_\infty + C_{\mathrm{P}}(2,\Omega,\Gamma)\, \alpha\, \|\operatorname{div} \mathbf{w}\|_3\Big), \end{aligned} \tag{3.12a}$$

$$\begin{aligned} \|Y^{-1}\|_{\mathcal{L}\left((L^2)^\#, L^2\right)} &\le 1 + \|\Pi\|_{\mathcal{L}(L^2)} \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)} C_{\mathrm{P}}(2,\Omega,\Gamma)\, |\Omega|^{\frac{1}{3}} \\ &\qquad \times \Big(\|\mathbf{w}\|_\infty + C_{\mathrm{P}}(2,\Omega,\Gamma)\, \alpha\, \|\operatorname{div} \mathbf{w}\|_3\Big). \end{aligned} \tag{3.12b}$$

Further assume that (2.20) *holds. If $\mathbf{w} \cdot \mathbf{n} \ge 0$ on $\partial\Omega \setminus \Gamma$ then X^{-1} is invertible, the inverse* *is given by* (3.10a)*, and*

$$\begin{aligned}\|X\|_{\mathcal{L}\left(L^2,(L^2)^\#\right)} &\le 1 + \|\Pi\|_{\mathcal{L}(L^2)} \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)} C_{\mathrm{P}}(2,\Omega,\Gamma)\,|\Omega|^{\frac{1}{3}} \\ &\quad \times \Big(\|\mathbf{w}\|_\infty + C_{\mathrm{P}}(2,\Omega,\Gamma)\,\alpha\,\|\operatorname{div}\mathbf{w}\|_3\Big) \\ &\quad \times \Big(1 - \big|\alpha - \tfrac{1}{2}\big|\, C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\Big)^{-1}.\end{aligned} \tag{3.13a}$$

If $\mathbf{w}\cdot\mathbf{n} \le 0$ *on* $\partial\Omega \setminus \Gamma$ *then* Y^{-1} *is invertible, the inverse is given by* (3.10b)*, and*

$$\begin{aligned}\|Y\|_{\mathcal{L}\left(L^2,(L^2)^\#\right)} &\le 1 + \|\Pi\|_{\mathcal{L}(L^2)} \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)} C_{\mathrm{P}}(2,\Omega,\Gamma)\,|\Omega|^{\frac{1}{3}} \\ &\quad \times \Big(\|\mathbf{w}\|_\infty + C_{\mathrm{P}}(2,\Omega,\Gamma)\,\alpha\,\|\operatorname{div}\mathbf{w}\|_3\Big) \\ &\quad \times \Big(1 - \big|\alpha - \tfrac{1}{2}\big|\, C_{\mathrm{P}}(2,\Omega,\Gamma)^2\,\|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\Big)^{-1}.\end{aligned} \tag{3.13b}$$

Proof. Under the appropriate conditions of Theorem 3.2 formulas (3.9) and (3.10) make sense and it remains to prove the estimates (3.12) and (3.13). Note that the definition (3.9a) can be rewritten as $(MX^{-1} - I) : (Q^h)^\# \to (Q^h)^\# : f \mapsto g$ such that $u \in W^h$ is given by

$$\int_\Omega \nabla u \cdot \nabla v = \langle f, \Pi^\dagger(\Pi\Pi^\dagger)^{-1} v\rangle_{(Q^h)^\#,Q^h} \qquad \text{for all } v \in W^h \tag{3.14}$$

and $g \in (Q^h)^\#$ is given by

$$\langle g, q\rangle_{(Q^h)^\#,Q^h} = \int_\Omega (\mathbf{w}\cdot\nabla u + \alpha \operatorname{div}\mathbf{w}\, u)\,\Pi q \qquad \text{for all } q \in Q^h. \tag{3.15}$$

Testing by u in (3.14) and by $M^{-1}g$ in (3.15), noticing that $\langle g, M^{-1}g\rangle_{(Q^h)^\#,Q^h} = \|g\|^2_{(L^2)^\#}$ whenever $g \in (Q^h)^\#$, using (3.8), Hölder's inequality, and the Sobolev-Poincaré inequality $\|u\|_6 \le C_{\mathrm{P}}(2,\Omega,\Gamma)\|\nabla u\|_2$ yields the estimate (3.12a). Estimate (3.13a) then follows in a similar way using formula (3.10a) and obtaining the ellipticity (2.4) for $A + K$ in the same way as in the proof of Lemma 2.1. Concerning the Y-case, estimates (3.12b), (3.13b) follow in the same way. □

Remark 3.4. *We used the assumption* (2.2) *concerning the* $W^{1,3+\epsilon}$*-regularity of the Laplacian solve, which is fulfilled when* Ω *is a creased Lipschitz domain, to get a priori bounds* (2.17) *and* (2.44) *on the infinite-dimensional PCD operator* $X^{-1}_{\alpha,\mathbf{w},\Gamma}$ *and its inverse* $X_{\alpha,\mathbf{w},\Gamma}$*, which are uniform in* $\|\mathbf{w}\|_6 + \alpha\|\operatorname{div}\mathbf{w}\|_2 \le C\|\mathbf{w}\|_{1,2}$*, thus avoiding the dependence on* $\|\mathbf{w}\|_\infty + \alpha\|\operatorname{div}\mathbf{w}\|_3$*. The same can be achieved in the discrete case if the* $W^{1,3+\epsilon}$*-regularity carries over to the discrete Laplacian; if the problem: for* $f \in W^h$ *find* $u \in W^h$ *such that*

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f\, v \qquad \textit{for all } v \in W^h$$

admits the estimate

$$\|\nabla u\|_{3+\epsilon} \le C(W^h)\|f\|_2, \tag{3.16}$$

with some $\epsilon \ge 0$ *if* $\alpha = 0$ *and* $\epsilon > 0$ *if* $\alpha \in (0,1]$*, then one would get*

$$\|X^{-1}\|_{\mathcal{L}\left((L^2)^\#,L^2\right)} \le C\Big(\Omega,\, \Gamma,\, C(W^h),\, \|\Pi\|_{\mathcal{L}(L^2)},\, \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)},\, \|\mathbf{w}\|_6,\, \alpha\|\operatorname{div}\mathbf{w}\|_2\Big), \tag{3.17a}$$

$$\|Y^{-1}\|_{\mathcal{L}\left((L^2)^\#,L^2\right)} \le C\Big(\Omega,\, \Gamma,\, C(W^h),\, \|\Pi\|_{\mathcal{L}(L^2)},\, \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)},\, \|\mathbf{w}\|_6,\, \alpha\|\operatorname{div}\mathbf{w}\|_2\Big), \tag{3.17b}$$

and, if further (2.20) *holds, then*

$$\begin{aligned}\|X\|_{\mathcal{L}\left(L^2,(L^2)^\#\right)} \le C\Big(&\Omega,\, \Gamma,\, C(W^h),\, \|\Pi\|_{\mathcal{L}(L^2)},\, \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)},\, \|\mathbf{w}\|_6,\, \alpha\|\operatorname{div}\mathbf{w}\|_2,\\ &\big(1-|\alpha-\tfrac{1}{2}|\, C_{\mathrm{P}}(2,\Omega,\Gamma)^2\, \|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\big)^{-1}\Big)\end{aligned} \tag{3.18a}$$

if $\mathbf{w}\cdot\mathbf{n} \ge 0$ *on* $\partial\Omega\setminus\Gamma$*, and*

$$\begin{aligned}\|Y\|_{\mathcal{L}\left(L^2,(L^2)^\#\right)} \le C\Big(&\Omega,\, \Gamma,\, C(W^h),\, \|\Pi\|_{\mathcal{L}(L^2)},\, \|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)},\, \|\mathbf{w}\|_6,\, \alpha\|\operatorname{div}\mathbf{w}\|_2,\\ &\big(1-|\alpha-\tfrac{1}{2}|\, C_{\mathrm{P}}(2,\Omega,\Gamma)^2\, \|\operatorname{div}\mathbf{w}\|_{\frac{3}{2}}\big)^{-1}\Big)\end{aligned} \tag{3.18b}$$

if $\mathbf{w}\cdot\mathbf{n} \le 0$ *on* $\partial\Omega\setminus\Gamma$*. The proof of this follows along the same lines as the proofs of Theorem 3.3, Lemma 2.5, and Theorem 2.15.*

Due to assumption (3.1)*, which we have deliberately taken, the regularity estimate* (3.16) *is always true for a fixed space* W^h *with some* $C(W^h) > 0$*. Of course, it would be desirable if the constant was bounded uniformly with refinement, which would render estimates* (3.17) *and* (3.18) *independent of mesh refinement as well. Nevertheless, it seems that a uniform estimate* (3.16) *is problematic even for quasi-uniform refinement; the standard technique, which uses an inverse estimate and convergence order* $h^{\frac{3}{2}+\epsilon}$ *in the* L^2*-norm, fails here because* $W^{\frac{3}{2}+\epsilon,2}$*-regularity of the Laplacian does not hold even in the Dirichlet case; see [11].*

Concerning the case of adaptive refinement, the refinement of pressure space Q^h *would have to ensure validity of* (3.16)*. Indeed, we will typically define* W^h *in terms of* Q^h *in the subsequent sections; we already know that* W^h *is related to* Q^h *by* (3.5)*.*

In other words, unsuitable or insufficient refinent of the pressure space can spoil the validity of the uniform estimates (3.17) *and* (3.18) *and thus the performance of the preconditioner in the worst case. On the other hand, the uniform estimates of Theorem 3.3, which depend on* $\|\mathbf{w}\|_\infty + \alpha \operatorname{div} \| \operatorname{div} \|_3$*, are valid for arbitrary refinement. For better understading, whether and/or when the dependence on* $\|\mathbf{w}\|_\infty + \alpha \operatorname{div} \| \operatorname{div} \|_3$ *or* $\|\mathbf{w}\|_6 + \alpha \operatorname{div} \| \operatorname{div} \|_2$ *is acceptable/appropriate, it is necessary to take into account the employed non-linear iteration scheme and its a priori estimates; we remark here once more that the underlying non-linear problem* (1.1) *does not admit a priori estimates in general cases, i.e., for large data, cf. the discussion in Section 1. Nevertheless, this is beyond the scope of this work.*

Remark 3.5. *The bounds on the Schur complement in Lemma 2.3 are true in the discrete case with* $\beta(\Omega,\mathrm{D})$ *replaced by the velocity-pressure discrete inf-sup constant, provided that the velocity-pressure pair discrete space is inf-sup stable; we did not treat any stabilized case in this study. This implies, together with the bounds* (3.12) *and* (3.13)*, or* (3.17) *and* (3.18)*, a spectral bound on the discrete preconditioned Schur complement analogous to* (2.47)*, with a modification of* (2.48) *which additionally depends on the discrete inf-sup constant, the norms* $\|\Pi\|_{\mathcal{L}(L^2)}$ *and* $\|\Pi^\dagger(\Pi\Pi^\dagger)^{-1}\|_{\mathcal{L}(L^2)}$*, and, in the case* $\|\mathbf{w}\|_6 + \alpha\|\operatorname{div}\mathbf{w}\|_2$*, the constant* $C(W^h)$ *of the discrete* $W^{1,3+\epsilon}$*-estimate* (3.16)*, all under the appropriate condition on the wind direction on* $\partial\Omega\setminus\Gamma$*.*

The result of Theorem 2.16, which relates GMRES convergence of the preconditioned saddle-point system to GMRES convergence of the preconditioned Schur complement with the lag L, stays true in the discrete case under technical conditions on the discretization which we do not discuss. The lag L, i.e., the length of the Jordan chain, is finite in the infinite-dimensional case and thus there is hope that it is uniformly bounded in the discrete case; nevertheless we did not prove this. On the other hand, the results of Sections 2.4 and 2.5 are not obviously transferable to the discrete case; indeed, the decay rates (2.70) *might be difficult to establish in the discrete case and would likely depend on appropriate approximation properties of the involved discrete spaces and the transfer operator* Π*.*

3.2 Case of continuous pressure discretizations

Now we consider the important case where Q^h is smooth enough such that it contains a subspace suitable for conforming discretizations of Laplacian and convection-diffusion; specifically, when $Q^h \subset W^{1,\infty}_\Gamma$, we can consider the choice of W^h given by

$$W^h = Q^h \cap W^{1,\infty}_\Gamma. \tag{3.19}$$

Then we can consider the transfer operator $\Pi : Q^h \to W^h$ given by

$$(\Pi q, w)_{W^h} = (q, w)_{Q^h} \qquad \text{for all } q \in Q^h,\ w \in W^h. \tag{3.20}$$

Note that this definition is only valid thanks to the fact that the right-hand side of (3.20) makes sense due to the inclusion $W^h \subset Q^h$, which is a consequence of (3.19). As an important example consider Q^h being a continuous finite element space; specifically, assuming Ω is a polytopic domain partitioned into simplicial conforming mesh $\mathcal{T}_h$, define, for integer $k \geq 1$,

$$\begin{aligned} \mathcal{P}_k(\mathcal{T}_h) &\coloneqq \{p : \Omega \to \mathbb{R},\ p_{|K} \text{ polynomial of total degree at most } k \text{ for all } K \in \mathcal{T}_h\}, \\ Q^h &\coloneqq \{q \in W^{1,\infty},\ q \in \mathcal{P}_k(\mathcal{T}_h)\}. \end{aligned} \tag{3.21}$$

Indeed, such Q^h is the space of continuous, piece-wise polynomials of degree at most k and W^h given by (3.19) is merely its subspace with zero boundary values on Γ.

Now we turn to the derivation of a specific form of X^{-1} and Y^{-1} for the case (3.19), (3.20). Relations (3.3) and (3.20) imply that

$$(q, \Pi^\dagger w)_{Q^h} = (q, w)_{Q^h} \qquad \text{for all } q \in Q^h,\ w \in W^h$$

which means that $\Pi^\dagger : W^h \to Q^h$ is merely the inclusion (identity) operator $W^h \subset Q^h$. Hence, in block form, with blocks corresponding to W^h and Q^h/W^h,

$$\Pi^\dagger = \begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{3.22}$$

Note that in the case (3.21) with the standard *dual basis* (also *nodes* or *degrees of freedom*), the splitting $W^h \oplus Q^h/W^h$ is identified with DOFs interior to $\overline{\Omega} \setminus \Gamma$ and Γ-boundary DOFs. Relations (3.22) and (3.4) imply

$$\Pi = (M^W)^{-1} \begin{pmatrix} M^Q_{11} & M^Q_{12} \end{pmatrix} \tag{3.23}$$

with M^Q_{ij} blocks corresponding to W^h and Q^h/W^h in

$$M^Q = \begin{pmatrix} M^Q_{11} & M^Q_{12} \\ M^Q_{21} & M^Q_{22} \end{pmatrix}.$$

Taking the special choice

$$M^W \coloneqq M^Q_{11} \tag{3.24}$$

yields

$$\Pi = \begin{pmatrix} I & (M^Q_{11})^{-1} M^Q_{12} \end{pmatrix}, \tag{3.25a}$$

$$\Pi^\dagger (\Pi \Pi^\dagger)^{-1} = \begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{3.25b}$$

Note that so far we have not fixed the inner product on Q^h. We argue that it makes sense to choose the L^2-inner product, i.e.,

$$(q_1, q_2)_{Q^h} = \langle M^Q q_1, q_2 \rangle_{(Q^h)^\# \times Q^h} = \int_\Omega q_1\, q_2 \qquad q_1, q_2 \in Q^h. \tag{3.26}$$

Indeed, it is easy to check that with this choice we have

$$\|P\|_{\mathcal{L}(L^2)} \leq 1, \qquad \|\Pi\|_{\mathcal{L}(L^2)} \leq 1, \qquad \|\Pi^\dagger\|_{\mathcal{L}(L^2)} \leq 1, \tag{3.27}$$

so that Π fulfills the conditions of Theorem 3.2 and (3.27) yields estimates (3.12) and (3.13) independent of mesh refinement. On the other hand, in the typical finite element settings (3.21) with $\{\phi_j\}_{j=1}^{N^Q}$ being the usual local basis functions of the space (3.21), one can use only the diagonal part of the L^2-inner product (3.26), i.e.,

$$(\phi_i, \phi_j)_{Q^h} = \langle M^Q \phi_i, \phi_j \rangle_{(Q^h)^\# \times Q^h} = \begin{cases} \int_\Omega \phi_i \, \phi_j & i = j \\ 0 & i \neq j \end{cases} \qquad i, j = 1, \dots, N^Q. \tag{3.28}$$

The advantage of this is that $M_{12}^Q = 0$; subtituting into (3.25) yields

$$\Pi = \begin{pmatrix} I & 0 \end{pmatrix}, \tag{3.29a}$$

$$\Pi^\dagger (\Pi \Pi^\dagger)^{-1} = \begin{pmatrix} I \\ 0 \end{pmatrix}, \tag{3.29b}$$

which is a particularly simple form of the transfer operators in (3.9). Under certain conditions, (3.28) is equivalent to (3.26) uniformly with mesh refinement, which allows one to recover L^2-estimates (3.27) also for the diagonal inner product (3.28); specifically, Wathen [63] shows that (3.26) and (3.28) are equivalent for Lagrange elements on simplices under *any* refinement and points out that the equivalence is also true on quads/hexes under shape-regular refinement.

Summarizing, with the choice of inner products (3.24) we have

$$X^{-1} = M^{-1} \left(\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \begin{pmatrix} I \\ M_{21}^Q (M_{11}^Q)^{-1} \end{pmatrix} K A^{-1} \begin{pmatrix} I & 0 \end{pmatrix} \right), \tag{3.30a}$$

$$Y^{-1} = \left(\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \begin{pmatrix} I \\ 0 \end{pmatrix} A^{-1} K^{\mathrm{R}} \begin{pmatrix} I & (M_{11}^Q)^{-1} M_{12}^Q \end{pmatrix} \right) M^{-1}, \tag{3.30b}$$

which is further equivalent to

$$X^{-1} = M^{-1} \left(\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \begin{pmatrix} I & 0 \\ M_{21}^Q (M_{11}^Q)^{-1} & 0 \end{pmatrix} \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \right), \tag{3.31a}$$

$$Y^{-1} = \left(\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} K^{\mathrm{R}} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} I & (M_{11}^Q)^{-1} M_{12}^Q \\ 0 & 0 \end{pmatrix} \right) M^{-1}. \tag{3.31b}$$

Note that with the choice of the diagonal inner product (3.28) the terms M_{12}^Q and M_{21}^Q in (3.30) and (3.31) vanish. Also note that (3.31) can be equivalently rewritten without the projection factors $\left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$ but it might be better to keep them in the case that an approximation in $\left(\begin{smallmatrix} A & 0 \\ 0 & I \end{smallmatrix}\right)^{-1}$ leads to loss of the projection property.

Consider operator $G : Q^h \to Q^h$ given by

$$G = \begin{pmatrix} I & (M_{11}^Q)^{-1} M_{12}^Q \\ 0 & I \end{pmatrix}.$$

This is a similarity transformation which diagonalizes P and $I - P$, i.e.,

$$G^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} G = P, \qquad GPG^{-1} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

$$G^{-1} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} G = I - P, \qquad G(I - P)G^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}.$$

It is straightforward to check that (3.31) is equivalent to

$$X^{-1} = M^{-1} G^\# \begin{pmatrix} A + K & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} G^{-\#}, \tag{3.32a}$$

$$Y^{-1} = G^{-1} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} A + K^{\mathrm{R}} & 0 \\ 0 & I \end{pmatrix} G M^{-1}. \tag{3.32b}$$

In the case of the diagonal inner product (3.28), where $M^Q_{12} = 0$, the formulas (3.32) immediately reduce to

$$X^{-1}_{\operatorname{diag} L^2} = M^{-1} \begin{pmatrix} A+K & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1}, \tag{3.33a}$$

$$Y^{-1}_{\operatorname{diag} L^2} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} A+K^{\mathrm{R}} & 0 \\ 0 & I \end{pmatrix} M^{-1}. \tag{3.33b}$$

Note that it might be useful to expand

$$\begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1}, \qquad \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}$$

in (3.33) to the explicit identity in the case that the solve is approximated and recovery of the identity is not guaranteed, i.e., to prefer definition (3.31) with $M^Q_{12} = M^Q_{21} = 0$, so that the leading order term is preserved even when $\left(\begin{smallmatrix} A & 0 \\ 0 & I \end{smallmatrix}\right)^{-1}$ is eventually approximated.

Let us consider once more the case that M^Q is the full L^2-inner product (3.26). In that case all of the formulas (3.30), (3.31), and (3.32) are valid. But consider that now $M^Q = M$ and

$$\begin{aligned} \begin{pmatrix} I & (M^Q_{11})^{-1} M^Q_{12} \\ 0 & 0 \end{pmatrix} M^{-1} = \begin{pmatrix} (M^Q_{11})^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} M^Q_{11} & M^Q_{12} \\ M^Q_{21} & M^Q_{22} \end{pmatrix} \begin{pmatrix} M^Q_{11} & M^Q_{12} \\ M^Q_{21} & M^Q_{22} \end{pmatrix}^{-1} \\ = \begin{pmatrix} M^Q_{11} & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

which we can use to simplify (3.31) into

$$X^{-1}_{L^2} = M^{-1} + \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} M^Q_{11} & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.34a}$$

$$Y^{-1}_{L^2} = M^{-1} + \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} K^{\mathrm{R}} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} M^Q_{11} & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.34b}$$

Again, some of the projection factors $\left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$ could be removed but it is not clear if this is beneficial, considering that the solves in (3.34) might be approximated and the identity recovery in the Q^h/W^h block might be lost, whereas keeping the projection factors is a very cheap operation (zeroing Q^h/W^h entries of the intermediate vectors). Notice that the extra cost of using formulas (3.34), which use the full L^2-inner product (3.26), compared to formulas (3.33), which use the diagonal inner product (3.28), is essentially one extra solve, the L^2-projection to W^h: $\left(\begin{smallmatrix} M^Q_{11} & 0 \\ 0 & I \end{smallmatrix}\right)^{-1} \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

Formulas (3.30) (or equivalently (3.31), (3.32), and the special case (3.34) for inner product (3.26)) seem to be new and provide a very explicit description of how the boundary conditions should be incorporated. Formula (3.33a) seems to have appeared in the literature in the form of a verbal description; we will provide a detailed comparison with published results in Section 3.5.

3.3 Case of higher-order discontinuous pressure discretizations

In this section we will sketch the construction of a discrete PCD preconditioner in the case that the inclusion $Q^h \subset W^{1,\infty}_\Gamma$ does not hold but the definition (3.19) is still valid. We will see below that this does not apply to the lowest-order discontinuous Lagrange pressures, which is the case we will treat in the next section. To our knowledge, the discontinuous case has not received much attention in the literature. The only study we are aware of, which treats the case of discontinuous pressures is [22, pp. 368–370], which is essentially a finite difference construction. We are not sure if the present functional approach can cover this construction. On the other hand the material in the previous Section 3.2 seems to be a refinement of previously published approaches, as will be described in Section 3.5.

The case we consider in this section is when W^h can be again defined by (3.19). This is clearly not possible when Q^h consists of piecewise constant functions. In this case (3.19) implies that W^h consists of functions constant in Ω and further $W^h = \{0\}$ when $|\Gamma| > 0$. This is certainly undesirable. Nevertheless when Q^h consists of piecewise discontinuous higher-order polynomials the definition (3.19) yields a non-trivial W^h. In this case most of the exposition in Section 3.2 remains valid. But one must be more careful with its interpretation. Specifically, the splitting $Q^h = W^h \oplus Q^h/W^h$ does not have a straightforward meaning. In the context of Section 3.2 it was clearly a splitting into the degrees of freedom corresponding to the boundary Γ and the remaining degrees of freedom. When Q^h is discontinuous this is no longer the case and the decomposition depends on the chosen inner product M^Q. Because of the aforementioned reasons, we refrain from going into details and merely comment that the formulas of Section 3.2 can be given meaning with some extra Riesz liftings (typically a mass matrix solve or its approximation) implementing the inclusion $W^h \subset Q^h$ and the quotient Q^h/W^h, cf. (3.22), which would propagate further into the final formulas for X^{-1} and Y^{-1}.

3.4 Case of piecewise constant pressure discretizations

This section is motivated by the case of piecewise constant pressures, which does not allow one to define W^h as a subset of Q^h, as explained above, and the construction of Section 3.2 does not apply. But first we will proceed generally, with minimal assumptions.

Define $\Pi : Q^h \to W^h$ by

$$(\Pi q, w)_{W^h} = \int_\Omega q\, w \qquad \text{for all } q \in Q^h,\ w \in W^h.$$

Using (3.2b) we can express

$$\Pi = (M^W)^{-1} L \tag{3.35}$$

where $L : Q^h \to (W^h)^\#$ is given by $\langle Lq, w\rangle_{(W^h)^\#, W^h} = \int_\Omega q\, w$. From (3.35), (3.3), and (3.2) we immediately get $\Pi^\dagger = (M^Q)^{-1} L^\#$. Assume that L is surjective. Then by virtue of (3.35) and Lemma 3.1, the operator $\Pi\Pi^\dagger$ is invertible, and hence

$$\Pi^\dagger (\Pi\Pi^\dagger)^{-1} = (M^Q)^{-1} L^\# \big(L (M^Q)^{-1} L^\#\big)^{-1} M^W. \tag{3.36}$$

After choosing the spaces Q^h and W^h and the inner products M^Q and M^W, formulas (3.8), (3.35), (3.36), and (3.9) fully define the preconditioners X^{-1} and Y^{-1}, provided that the choice of Q^h and W^h renders Π given by (3.35) surjective.

Motivated by the desire to obtain X^{-1} and Y^{-1} bounded in $\mathcal{L}(L^2)$, one would require M^Q to be the L^2-inner product in Q^h given by formula (3.26), or at least some equivalent inner product, e.g., (3.28), which is equivalent under certain circumstances. A pitfall is the term $L(M^Q)^{-1}L^\#$ in (3.36), which could be dense and thus infeasible to be inverted. Nevertheless, if Q^h is a space of discontinuous, piecewise polynomials, then with an appropriate local basis, e.g., the classical discontinuous Lagrange basis, M^Q is block-diagonal. In such case $L(M^Q)^{-1}L^\#$ is still sparse. Thus this construction makes sense for discontinuous pressures.

In the sequel we will consider the case of piecewise constant pressures. Assume Ω is polytopic, partitioned into a simplicial conforming mesh $\mathcal{T}_h$, which comprises of cells $K \in \mathcal{T}_h$. We assume that

$$Q^h = \{q \in L^2,\ q_{|K} = \text{const } \forall K \in \mathcal{T}_h\}, \tag{3.37a}$$

$$W^h = \{w \in W^{1,\infty}_\Gamma,\ w_{|K} \in \mathcal{P}_1(K)\ \forall K \in \mathcal{T}_h\}. \tag{3.37b}$$

The first concern is whether the necessary condition (3.5) holds for this choice. To provide the answer we first prove the following lemma.

Lemma 3.6. *Let $\Omega \subset \mathbb{R}^d$ be a bounded polytopic domain and $\mathcal{T}_h$ be a conforming simplicial triangulation of Ω. Then the number of vertices in $\mathcal{T}_h$ is less than or equal to the number of cells in $\mathcal{T}_h$ plus d.*

Proof. Removing any cell connected by a facet to some other cell removes at most one vertex. By virtue of the assumption that Ω is connected, the removal can be done on the boundary, so that the triangulations stays connected after the removal. Such removals can be repeated until one cell is left. The remaining cell has $d+1$ vertices. □

The validity of condition (3.5) is now established as a simple consequence of the lemma.

Corollary 3.7. *Let the conditions of Lemma 3.6 be fulfilled. Assume that $\Gamma \subset \partial\Omega$ satisfies $|\Gamma| > 0$. Then the spaces Q^h and W^h given by* (3.37) *satisfy condition* (3.5)*.*

On the other hand, very simple counterexamples show that the necessary condition (3.5) does not hold for (3.37) with $\mathcal{P}_1$ replaced by $\mathcal{P}_2$. We gain further confidence in the choice (3.37), (3.35) by showing that Π is surjective and thus by virtue of Lemma 3.1 (3.36) is well defined.

Theorem 3.8. *Let $\Omega \subset \mathbb{R}^d$ be a bounded polytopic domain and $\mathcal{T}_h$ be a conforming simplicial triangulation of Ω. Further let $\Gamma \subset \partial\Omega$ be such that $|\Gamma| > 0$. Then for Q^h and W^h given by* (3.37)*, Π given by* (3.35) *is surjective, $\Pi\Pi^\dagger$ is invertible, and thus* (3.36) *is well defined.*

Proof. By virtue of Lemma 3.1, the surjectivity of Π and invertibility of $\Pi\Pi^\dagger$ are equivalent to the injectivity of $\Pi^\#$, which is in turn equivalent to the injectivity of $L^\#$.

Choose $w \in W^h$ such that $L^\# w = 0$. Hence

$$0 = \int_\Omega \phi_i\, w = \int_{K_i} w \qquad \text{for all } i = 1, 2, \dots, N^Q. \tag{3.38}$$

On any cell K adjacent to a facet intersecting Γ, $w_{|K}$ is affine, vanishing on such facet, and, by virtue of (3.38), has zero mean over K. Hence $w_{|K} = 0$. By virtue of the assumption that Ω is connected, one can iterate this argument throughout the whole mesh in order to arrive at $w = 0$. Hence $L^\#$ is injective and the proof is finished. □

Theorem 3.8 ensures, in particular, that Π is surjective and that Theorem 3.2, which guarantees invertibility of X^{-1} or Y^{-1} under the appropriate conditions, can be invoked.

Finally, we provide explicit matrix entries of the operators (3.35) and (3.36) for the choice of the spaces (3.37) and particular choices of M^Q and M^W. Consider the standard basis of Q^h: $\phi^i_{|K_j} = \delta_{ij}$, $i, j = 1, 2, \dots, N^Q$. Then $M^Q = M$ is diagonal with entries $|K_i|$, $i = 1, 2, \dots, N^Q$. Now consider a choice of the inner product on W^h:

$$(w_1, w_2)_{W^h} = \langle M^W w_1, w_2 \rangle_{(W^h)^\# \times W^h} = \int_\Omega w_1\, w_2 \qquad w_1, w_2 \in W^h. \tag{3.39}$$

Hence we have

$$M^W_{ij} = \int_\Omega \psi_i\, \psi_j \qquad i, j = 1, 2, \dots, N^W, \tag{3.40a}$$

$$L_{ij} = \int_\Omega \psi_i\, \phi_j \qquad i = 1, 2, \dots, N^W,\ j = 1, 2, \dots, N^Q, \tag{3.40b}$$

which implies that

$$\begin{aligned}\big(L(M^Q)^{-1}L^\#\big)_{ij} &= \sum_{k=1}^{N^Q} \frac{\int_\Omega \psi_i\, \phi_k \int_\Omega \psi_j\, \phi_k}{\int_\Omega |\phi_k|^2} = \sum_{K \in \omega_i \cap \omega_j} \frac{\int_K \psi_i \int_K \psi_j}{|K|} \\ &= \frac{|\omega_i \cap \omega_j|}{(d+1)^2} \qquad i, j = 1, 2, \dots, N^W,\end{aligned} \tag{3.40c}$$

$$\big((M^Q)^{-1}L^\#\big)_{ij} = \frac{\int_\Omega \phi_i\, \psi_j}{\int_\Omega |\phi_i|^2} = \begin{cases} \frac{1}{d+1} & \text{if } K_i \subset \omega_j, \\ 0 & \text{otherwise} \end{cases} \qquad \begin{cases} i = 1, 2, \dots, N^Q, \\ j = 1, 2, \dots, N^W, \end{cases} \tag{3.40d}$$

where ω_j is the patch of the cells adjacent to vertex j. Formulas (3.35), (3.36), and (3.40) give a precise definition of the transfer operators Π and $\Pi^\dagger(\Pi\Pi^\dagger)^{-1}$ which appear in the definition of X^{-1} and Y^{-1}; see (3.9). Notice that the two extra solves are necessary due to the transfer operators, namely a solve with (3.40a) and a solve with (3.40c). Note that the first solve can be avoided by approximating M^W by its diagonal. This is an inner product which is (uniformly with mesh refinement) equivalent to (3.39); see Wathen [63]. Note also that it is desirable to use the same choice (approximation) of M^W in (3.35) and (3.36) for the validity of Theorem 3.2.

3.5 Historical remarks

There has been a consensus for a long time that natural boundary conditions for the Laplacian solve and convection-diffusion operator in the preconditioner are appropriate in the case of enclosed flows, i.e., when $\mathrm{D} = \partial\Omega$. In our framework this corresponds to $\Gamma = \emptyset$ although we did not treat this case as it requires special care since the pressure is determined up to a constant. Nevertheless this is well understood and it would be possible to incorporate it into our analysis for the price of a few more technical difficulties and if-thens.

On the other hand, it has been clear from the beginning that for inflow-outflow problems the treatment of boundary conditions in the preconditioner must be different in order to preserve good performance of the method. From today's perspective it seems that many contradictory, or at least vague, conclusions have been reached and published. Even the recent monograph by Elman, Silvester, and Wathen [22], together with its supporting code IFISS [58], a gives rather ambiguous account of the issue.

In this historical exposition we will focus only on the case of a continuous pressure space Q^h, in which we can define W^h by (3.19). Motivated by the enclosed flow case, which we omit, it is natural to consider

$$X^{-1} = M^{-1}\hat{F}\hat{A}^{-1}, \tag{3.41a}$$

$$Y^{-1} = \hat{A}^{-1}\hat{F}^{\mathrm{R}}M^{-1} \tag{3.41b}$$

with some $F, F^{\mathrm{R}}, A : Q^h \to (Q^h)^{\#}$ and (3.8a). Note that the cases (3.41a) and (3.41b) were not considered together until this study. In the enclosed flow case, it is well justified to choose

$$\langle Mp, q\rangle = \int_\Omega p\, q \qquad p, q \in Q^h, \tag{3.42a}$$

$$\langle \hat{A}p, q\rangle = \int_\Omega \nabla p \cdot \nabla q \qquad p, q \in Q^h, \tag{3.42b}$$

$$\langle \hat{K}p, q\rangle = \int_\Omega \mathbf{w} \cdot \nabla p\, q + \alpha \operatorname{div} \mathbf{w}\, p\, q \qquad p, q \in Q^h, \tag{3.42c}$$

$$\langle \hat{K}^{\mathrm{R}}p, q\rangle = \int_\Omega \mathbf{w} \cdot \nabla p\, q + \alpha \operatorname{div} \mathbf{w}\, p\, q - \int_{\partial\Omega\setminus\Gamma} \mathbf{w} \cdot \mathbf{n}\, p\, q \qquad p, q \in Q^h, \tag{3.42d}$$

$$\hat{F} = \hat{A} + \hat{K}, \qquad \hat{F}^{\mathrm{R}} = \hat{A} + \hat{K}^{\mathrm{R}} \tag{3.42e}$$

up to the technicality of dealing with constants.[7] Now, in the inflow-outflow case it has been demonstrated above that $\hat{A}$, $\hat{F}$, and $\hat{F}^{\mathrm{R}}$ have to be somehow modified to account for the boundary conditions in W^h. The necessity of this has been clear from the first studies of the preconditioner in view of experience with preconditioner performance. The approach of most studies consists of two steps:

1. assemble (3.42),
2. modify the action of (a subset of) M, $\hat{A}$, $\hat{F}$, and $\hat{F}^{\mathrm{R}}$ on the boundary (in a more or less ad hoc way) before plugging them into (3.41).

One of the first accounts of this issue states:

> In the case of a boundary segment with standard outflow boundary conditions, the Schur complement S (and its preconditioner $[X]$) must be defined with Dirichlet data for the pressure on that part of the boundary in order to ensure that the preconditioning operator is elliptic over the pressure solution space. (Elman, Silvester, and Wathen [20, p. 668])

[7] In the enclosed flow case, contrary to inflow-outflow problems, the pressure is determined up to a constant and it is natural to look for a pressure in a space isomorphic to $L^2/\mathbb{R}$, for example $\{q \in L^2, \int_\Omega q = 0\}$. Constants can be factored out from Q^h in the same or similar way. On the other hand it is convenient from an implementation standpoint to keep the full Q^h and use solvers which suitably handle the nontrivial kernel in consistent systems; Elman, Silvester, and Wathen [22, section 9.3.5] describe this paradigm with the phrase "singular systems are not a problem". Hence, for the case of enclosed flows, $\hat{A}^{-1}$ in (3.41) is to be understood as the solution operator for $\hat{A} : Q^h/\mathbb{R} \to (Q^h/\mathbb{R})^{\#}$.

This does not seem to provide a sufficiently specific method for defining X^{-1}. Moreover it is also very misleading, if not incorrect; the Schur complement, at least for inf-sup stable conforming discretizations, is a bijection $S : Q^h \to (Q^h)^\#$ and it does not seem to feature the aforementioned Dirichlet data. On the other hand, a later monograph by the same authors is more specific:

> [F]or the reduced problem [$\mathbf{w} \cdot \nabla u = f$] occurring in the limiting case of pure convection, the solution is determined by specified Dirichlet boundary conditions on the inflow boundary, where [$\mathbf{w} \cdot \mathbf{n} < 0$]. This suggests using Dirichlet boundary conditions along inflow boundaries to define [$\hat{F}$]. This means that the rows and columns of [$\hat{F}$] corresponding to pressure nodes on an inflow boundary are treated as though they are associated with Dirichlet boundary conditions. At nodes on other (characteristic or outflow) components of $\partial\Omega$, the entries of [$\hat{F}$] are defined by [do-nothing condition]. [$\hat{A}$] is defined in an analogous manner so that [$\hat{F}$] an [$\hat{A}$] are derived from consistent boundary conditions.
>
> We note that this discussion only concerns the definition of the algebraic operators, [$\hat{F}$] and [$\hat{A}$]. That is, the only place where boundary conditions have any impact is on the *definition* of the preconditioning operator. In particular, there are no boundary conditions imposed on the discrete pressures, no values of Dirichlet conditions to determine, and there is no right-hand side that is affected by these boundary node modifications. (Elman, Silvester, and Wathen [21, pp. 348–349])

Perhaps "the rows and columns [...] corresponding to pressure nodes on an inflow boundary are treated as though they are associated with Dirichlet boundary conditions" was meant to express that

$$\hat{A} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}, \tag{3.43}$$

$$\hat{F} = \begin{pmatrix} A + K & 0 \\ 0 & I \end{pmatrix}, \quad \hat{F}^{\mathrm{R}} = \begin{pmatrix} A + K^{\mathrm{R}} & 0 \\ 0 & I \end{pmatrix} \tag{3.44}$$

with (3.8b), (3.8c), and (3.8d), which leads exactly to formulas (3.33). Even more intersting is the second paragraph of the above quote; one can guess that "no right-hand side [...] is affected by these boundary node modifications" means that (3.43) is devised to be used instead of

$$\hat{A}^{-1} := \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.45}$$

Indeed, (3.45) is a solution operator for the problem: for a given $f \in (Q^h)^\#$ find $u \in W^h = Q^h \cap W^{1,\infty}_\Gamma$ such that

$$\int_\Omega \nabla u \cdot \nabla v = \langle f, v \rangle \qquad \text{for all } v \in W^h. \tag{3.46}$$

Notice once more that (3.45) really "implements" the boundary-value problem (3.46). It is crucial to observe that using (3.44) and (3.45) renders both (3.41) singular, simply because (3.45) is singular.[8] This demonstrates how much potential confusion can occur due to the contradiction between (3.43) and (3.45). Perhaps Olshanskii and Vassilevski [52, paragraph under (2.7)] thought that the formula (3.45) should be used rather than (3.43). This is manifested in the encountered singularity of the precondioner which they circumvent by enforcing the Dirichlet boundary condition on a fictitious boundary slightly outside of Ω introduced along Γ:

[8] Indeed, the right-hand side of (3.45) is singular when taken as an operator $(Q^h)^\# \to Q^h$, but the motivation for calling it an inverse operator is that it can be viewed as a solution operator to (3.46) in the sense that

$$\hat{A}^{-1} = (Q^h)^\# \hookrightarrow (W^h)^\# \overset{A^{-1}}{\to} W^h \hookrightarrow Q^h$$

with $\left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$ representing the inclusion operator $(Q^h)^\# \hookrightarrow (W^h)^\#$.

> [I]n the continuous counterpart of the preconditioner ($[X^{-1}_{\alpha,\mathbf{w},\Gamma} = F_{\alpha,\mathbf{w},\Gamma}A^{-1}_{\Gamma}]$) one has to prescribe boundary conditions *only* for the Poisson problem solution operator $[A^{-1}_{\Gamma}]$, while in the discrete case some boundary conditions are involved in the definition of *both* matrices $[\hat{F}]$ and $[\hat{A}]$. [...] Furthermore, from the implementation standpoint, Dirichlet boundary conditions for $[\hat{F}]$ may not be imposed on the nodes at $[\Gamma]$ since these nodes have to contribute to the set of pressure degrees of freedom. For this reason one introduces outside Ω a fictitious one-cell layer attached to $[\Gamma]$. Dirichlet boundary conditions are assigned at layer nodes not belonging to $[\Gamma]$. (Olshanskii and Vassilevski [52, p. 2691])

> Dirichlet boundary conditions may not be imposed on the boundary nodes, since these nodes contribute to the set of pressure degrees of freedom. The Dirichlet condition is imposed on fictitious boundary nodes of an h-extension of the original mesh. Therefore, the actual boundary nodes are considered as interior in the extended mesh. This may be implemented in two ways. For a rectangular mesh we simply copy matrix entries for interior nodes to matrix entries for actual boundary nodes. However, for a general mesh one has to generate the fictitious mesh layer by reflecting the close-to-boundary layer of cells with respect to the actual boundary. (Olshanskii and Vassilevski [52, pp. 2700–2701])

Perhaps under the influence of Olshanskii and Vassilevski [52], the second edition of Elman, Silvester, and Wathen [22] omits the verbal description which advocates (3.43) over (3.45) in the first edition [21, pp. 348–349]. In fact no details concerning this issue appear in the second edition while the reference implementation, the Matlab package IFISS, uses the technique of Olshanskii and Vassilevski [52], which is obvious from the comments

> file `navier_flow/fpzsetup_q2p1.m`:
>
> ```
> % Dirichlet b.c. for Ap and Fp at outflow: add effect of ``ghost points'' to the
> % right of the boundary back to the diagonal
> ```
>
> file `navier_flow/fpzsetup_q1.m`:
>
> ```
> % Dirichlet conditions for Ap and Fp at outflow boundary: mimic finite differences
> [...]
> % Augment diagonal with values from ``ghost elements'' outside outflow boundary
> ```
>
> (IFISS package version 3.3 [58])

We want to point out that our description of the discrete PCD operator, at least in the case of continuous pressures, see Section 3.2, leaves no space for confusion. All details concerning the boundary conditions are explicitly described by the formulas of the preceding sections, in contrast to verbal descriptions prevalent in previous studies. In fact our definitions of the PCD operators, (3.9) in the general case, (3.30)–(3.34), or (3.35), (3.36) in certain specific cases, are constructed such that the invertibility of the operators is ensured by Theorem 3.2. Precisely, provided that (2.20) holds, X^{-1} is invertible if $\mathbf{w} \cdot \mathbf{n} \geq 0$ on $\partial\Omega \setminus \Gamma$ and Y^{-1} is invertible if $\mathbf{w} \cdot \mathbf{n} \leq 0$ on $\partial\Omega \setminus \Gamma$. This corresponds to the choice of Γ being the inflow boundary for the X-variant as in [21, pp. 348–349], and [52, paragraph under (2.7)] and Γ chosen as the outflow boudary for the Y-variant as in [22, second paragraph on p. 372]. Furthermore, a priori bounds uniform in certain norms of the data are provided by Theorem 3.3 and Remark 3.4.

A completely different approach is chosen by Deuring [15], who for inflow-outflow problems substitutes the Dirichlet condition on Γ (2.1b) with the Robin condition

$$\frac{\partial r}{\partial \mathbf{n}} = -\kappa r \text{ on } \Gamma$$

using the special choice $\kappa = 1$. The borderline case $\kappa = 0$, the Neumann condition, is known to perform poorly for inflow-outflow problems. The limiting choice $\kappa \to \infty$, which corresponds to the Dirichlet condition on Γ, has had the most attention in the literature but suffers from the problems described above in this section. Using $\kappa < \infty$ can be seen as an attempt to circumvent these problems while it is indeed unclear why $\kappa = 1$ should be preferable choice. We would like to point out that [15] is a very careful and precise study, although it relies heavily on discretization properties like approximation properties (with quasi-uniform refinements), inverse estimates, etc., thus making it less general than our functional analytic approach of Section 2.

Next we would like to comment on eigenvalue bounds for the preconditioned Schur complement which appeared in the literature. To our knowledge, this work is the first study providing analysis in the function spaces, which is thus agnostic to discretizations. In the published literature there appear spectral bounds for the discrete operator which are analogous to (2.47), (2.48). In light of the aforementioned confusion regarding the boundary conditions, it is hard to imagine that the published spectral analyses apply to the inflow-outflow case considering that even the definition of the PCD operator is unclear. But even if we ignore this issue and focus on the enclosed flow case, $\Gamma = \emptyset$, $Q^h \subset \{q \in L^2, \int_\Omega q = 0\}$, a special case we do not treat in this study, the validity of published eigenvalue bounds for (3.41a) or (3.41b) with (3.42) (see [43, Corollary 9A], [18, Theorem 2.1], [42, p. 2046], [21, Theorem 8.5], [52, Theorem 3.2], [22, Theorem 9.9]) is still questionable. All these results boil down to [43, Theorem 7] which assumes the validity of $W^{2,2}$-estimates for the Laplacian and convection-diffusion, i.e., that $\hat{A}^{-1}, \hat{F}^{-1} \in \mathcal{L}(L^2, W^{2,2})$ uniformly with mesh refinement. This cannot be justified on general corner domains. Moreover, the bound is derived under the assumption $\operatorname{div} \mathbf{w} = 0$ and depends on $\|\mathbf{w}\|_\infty$, which are both very restrictive assumptions; cf. Lemma 2.7, Lemma 2.5. Apart of that, the result of Loghin [43] is very difficult to grasp because it works with and provides estimates in ℓ^2 matrix norms, which rely heavily on specific discretization properties (inverse estimates, approximation properties) that typically hold under quasi-uniform refinements and are difficult to transfer into general, possibly adaptive, refinements. We circumvent this by starting our analysis purely in the functional setting of Section 2, also called the *operator preconditioning* approach, and by working in problem-dictated norms rather than ℓ^2; in Section 3.1 we transfer the results of the analysis to some discrete cases.

At this point we would like to note that some studies, e.g., [22, section 9.3.4], give the impression that spectral bounds of the preconditioned Schur complement are useful for bounding the GMRES convergence rate. Consider that for diagonalizable operator $T = V^{-1} \operatorname{diag}(\{\lambda_i\}_i) V$ and polynomial p it holds that

$$\|p(T)\| \le \|V\| \, \|V^{-1}\| \max_{t \in \{\lambda_i\}_i} |p(t)|.$$

Furthermore, when T is normal then V is unitary so that $\|V\| = \|V^{-1}\| = 1$. Nevertheless there does not seem to be any evidence of normality or diagonalizability of the preconditioned Schur complement in general geometry. It is important to recall that spectrum does not give, in general, any valid bound on the convergence of GMRES as noticed by Greenbaum, Pták, and Strakoš [26].

Appendix A Spectrum of bounded linear operators

In this section we restate some classical textbook matter concerning spectra of bounded linear operators as well as recent results describing the rate of accumulation of eigenvalues at essential spectrum for certain classes of operators.

First we state Banach's bounded inverse theorem, which is a corollary of the open mapping theorem. For a proof see [17, Theorem II.2.2] or [64, Corollary on p. 77].

Theorem A.1 (Bounded inverse theorem). *Let V be a Banach space and let T be a bounded linear operator on V. If T is bijective then the inverse operator $T^{-1} : V \to V$ is bounded.*

We continue with a characterization of invertibility of a bounded linear operator.

Lemma A.2 (Characterization of invertibility). *Let V be a Banach space and T a bounded linear operator on V. The inverse operator $T^{-1} \in \mathcal{L}(V)$ exists if and only if the following conditions are met:*

(i) T is bounded from below, i.e., there exists $c > 0$ such that

$$\|T\,x\|_V \geq c\,\|x\|_V \qquad \text{for all } x \in V, \tag{A.1}$$

(ii) the range of T is dense, i.e.,

$$\overline{\{T\,x:\ x \in V\}}^V = V. \tag{A.2}$$

Proof. First assume that conditions (i), (ii) hold. Condition (i) implies that T is injective and the range of T is closed, which together with (ii) implies that T is surjective. Hence T^{-1} exists and by (A.1) is bounded.

On the other hand if (i) does not hold, there exists $\{x_k\}_{k=1}^{\infty} \subset V$ with $\|x_k\|_V = 1$ and $Tx_k \to 0$. If T^{-1} existed and was bounded, then it would hold that $x_k = T^{-1}Tx_k \to 0$, which is a contradiction. Hence (i) is necessary. Condition (ii) is obviously necessary. □

The spectrum $\sigma(T)$ of a bounded operator T on a Banach space is a subset of the complex plane containing those λ such that $T - \lambda I$ is not continuously invertible, or equivalently by Theorem A.1, $T - \lambda I$ is not invertible. By the characterization of Lemma A.2 the spectrum can be broken down into two, not necessarily disjoint, parts:

1. the approximate point spectrum $\sigma_{\mathrm{ap}}(T)$ consisting of those $\lambda \in \mathbb{C}$ such that $T - \lambda I$ is not bounded from below; this is equivalently characterized by the following: there exists a sequence $\{x_k\} \subset V$, $\|x_k\|_V = 1$ such that

$$\|(T - \lambda I)x_k\|_V \to 0; \tag{A.3}$$

   if λ is an eigenvalue, i.e., there exists $x \in V$, $\|x\|_V = 1$ such that $Tx - \lambda x = 0$, then one can choose the sequence in (A.3) as $x_k := x$; henceforth we define the point spectrum $\sigma_{\mathrm{p}}(T) \subset \sigma_{\mathrm{ap}}(T)$ as the set of eigenvalues;

2. the compression spectrum $\sigma_{\mathrm{cp}}(T)$ consisting of those $\lambda \in \mathbb{C}$ such that the closure of the range of $T - \lambda I$ is a proper subset of V.

We proceed by bounding the approximate point spectrum by the operator norm.

Lemma A.3. *Let V be a Banach space. Let the operator $T \in \mathcal{L}(V)$ have an inverse $T^{-1} \in \mathcal{L}(V)$. Then the approximate point spectrum of the operator T is contained in the set*

$$\left\{\lambda \in \mathbb{C} : \|T^{-1}\|_{\mathcal{L}(V)}^{-1} \leq |\lambda| \leq \|T\|_{\mathcal{L}(V)}\right\}. \tag{A.4}$$

Proof. Let $\lambda \in \mathbb{C}$ be from the approximate point spectrum of T. Then by characterization (A.3) there exists a unit sequence x_k such that $Tx_k - \lambda x_k$ goes to zero in V. By the triangle inequality, considering that $\|x_k\|_V = 1$,

$$0 \leftarrow \|Tx_k - \lambda x_k\|_V \geq |\,\|Tx_k\|_V - |\lambda|\,|\,,$$

hence $\|Tx_k\|_V \to |\lambda|$. But $\|Tx_k\|_V$ is bounded from above by $\|T\|_{\mathcal{L}(V)}$ and hence is $|\lambda|$.

On the other hand, the sequence $\|x_k - \lambda T^{-1}x_k\|_V$ goes to zero because $\|x_k - \lambda T^{-1}x_k\|_V \leq \|T^{-1}\|_{\mathcal{L}(V)}\|Tx_k - \lambda x_k\|_V \to 0$. Hence $|\lambda|^{-1} \leftarrow \|T^{-1}x_k\|_V \leq \|T^{-1}\|_{\mathcal{L}(V)}$ and therefore $|\lambda|^{-1} \leq \|T^{-1}\|_{\mathcal{L}(V)}$. □

Now we show inclusion of the compression spectrum in the point spectrum of the adjoint.

Lemma A.4. *Let V be a Banach space and T a bounded linear operator on V. Let $V^{\#}$ denote the topological dual of V. Let $T^{\#} : V^{\#} \to V^{\#}$ be an adjoint of T defined by*

$$\left\langle \phi, Tx \right\rangle_{V^{\#},V} = \left\langle T^{\#}\phi, x \right\rangle_{V^{\#},V} \qquad \text{for all } x \in V \text{ and } \phi \in V^{\#}. \tag{A.5}$$

Then

$$\|T^{\#}\|_{\mathcal{L}(V^{\#})} = \|T\|_{\mathcal{L}(V)} \tag{A.6}$$

and $\sigma_{\mathrm{cp}}(T) \subset \sigma_{\mathrm{p}}(T^{\#})$.

Proof. The adjoint $T^\#$ is bounded and its norm is

$$\|T\|_{\mathcal{L}(V)} = \sup_{x\in V,\, \phi\in V^\#} \frac{\big\langle \phi, Tx\big\rangle_{V^\#,V}}{\|\phi\|_{V^\#}\,\|x\|_V} = \sup_{x\in V,\, \phi\in V^\#} \frac{\big\langle T^\#\phi, x\big\rangle_{V^\#,V}}{\|\phi\|_{V^\#}\,\|x\|_V} = \|T^\#\|_{\mathcal{L}(V^\#)}, \tag{A.7}$$

which proves (A.6).

Let $\lambda \in \sigma_{\mathrm{cp}}(T)$. Hence by definition $\overline{(T-\lambda I)V}^V$ is a proper subset of V. By the Hahn-Banach theorem there exists non-zero $\phi \in V^\#$ vanishing on $\overline{(T-\lambda I)V}^V$. Hence for all $x \in V$ it holds $0 = \langle \phi, (T-\lambda I)x\rangle_{V^\#,V} = \langle (T^\# - \lambda I)\phi, x\rangle_{V^\#,V}$. So λ is an eigenvalue of $T^\#$. □

Now we are in the position to bound the spectrum by the norm from above and by the inverse norm from below.

Theorem A.5. *Let V be a Banach space and T be a bounded linear operator on V. Then the spectrum of T fulfills the inclusion*

$$\sigma(T) \subset \Big\{\lambda \in \mathbb{C} : \|T^{-1}\|^{-1}_{\mathcal{L}(V)} \le |\lambda| \le \|T\|_{\mathcal{L}(V)}\Big\}, \tag{A.8}$$

with the convention $\|T^{-1}\|^{-1}_{\mathcal{L}(V)} = 0$ whenever T is not invertible.

Proof. First assume that T is invertible. By Lemma A.3 we already have $\|T^{-1}\|^{-1}_{\mathcal{L}(V)} \le |\lambda| \le \|T\|_{\mathcal{L}(V)}$ for $\lambda \in \sigma_{\mathrm{ap}}(T)$. By virtue of (A.6) the norms of T and $T^\#$ are equal, as well as the norms of T^{-1} and $T^{-\#}$, the adjoint of T^{-1} defined by (A.5). Now let $\lambda \in \sigma_{\mathrm{cp}}(T)$. By Lemma A.4 $\lambda \in \sigma_{\mathrm{p}}(T^\#)$ and also, by definition, $\lambda \in \sigma_{\mathrm{ap}}(T^\#)$. Now we can apply Lemma A.3 to $T^\#$ to conclude, with the help of (A.7), that $\|T^{-1}\|^{-1}_{\mathcal{L}(V)} = \|T^{-\#}\|^{-1}_{\mathcal{L}(V^\#)} \le |\lambda| \le \|T^\#\|_{\mathcal{L}(V^\#)} = \|T\|_{\mathcal{L}(V)}$. By Lemma A.2 we have that $\sigma_{\mathrm{ap}}(T) \cup \sigma_{\mathrm{cp}}(T) = \sigma(T)$. The case of T not invertible is a simple modification. □

Remark A.6. *In fact there is an elementary proof of Theorem A.5, in contrast to the given proof, which invokes the Hahn-Banach theorem through Lemma A.4, and the open mapping theorem through Theorem A.1. Indeed, if $|\lambda| > \|T\|_{\mathcal{L}(V)}$ then $(I - T/\lambda)^{-1}$ can be expressed by the Neumann series. Similarly, for $|\lambda| < \|T^{-1}\|^{-1}_{\mathcal{L}(V)}$ the Neumann series of $(I - \lambda T^{-1})^{-1} = T(T-\lambda I)^{-1}$ exists.*

Now, for a linear operator T on a complex Hilbert space H with inner product $(\cdot,\cdot)_H$ we define the numerical range of the operator as a subset of the complex plane given by

$$\mathrm{Num}(T) = \big\{(Tx, x)_H,\ \|x\|_H = 1\big\}.$$

The following statement is taken from Davies [14, Theorem 9.3.1].

Theorem A.7. *Let H be a complex Hilbert space and T be a bounded linear operator on H. Then* $\mathrm{Num}(T)$ *is a convex set and*

$$\sigma(T) \subset \overline{\mathrm{Num}}(T) \subset \big\{z \in \mathbb{C},\ |z| \le \|T\|\big\}$$

where $\overline{\mathrm{Num}}(T)$ *is the closure of* $\mathrm{Num}(T)$.

Proof. For the proof of the convexity of $\mathrm{Num}(T)$, which is known as the Toeplitz-Hausdorff theorem, see [14, Theorem 9.3.1], or [27].

The right-hand side inclusion is trivial. For the left-hand side inclusion, consider first $\lambda \in \sigma_{\mathrm{ap}}(T)$, i.e., there is a sequence $\{x_k\}_{k=1}^\infty \subset H$ with $\|x_k\|_H = 1$ and $\|(T-\lambda I)x_k\|_H \to 0$. That implies $|(Tx_k, x_k)_H - \lambda| \to 0$, which means that $\lambda \in \overline{\mathrm{Num}}(T)$. On the other hand, if $\lambda \in \sigma_{\mathrm{cp}}(T)$, then by Lemma A.4 there is $\phi \in H^\#$ with $\|\phi\|_{H^\#} = 1$ such that $T^\#\phi = \lambda\phi$. Equivalently, by the Riesz representation theorem, there is $x \in H$ with $\|x\|_H = 1$ and $T^\dagger x = \overline{\lambda} x$, where $T^\dagger : H \to H$ is a Hilbert adjoint of T, which is given by

$$(Ty, z)_H = (y, T^\dagger z)_H \qquad \text{for all } y, z \in H.$$

That implies $(Tx, x)_H = (x, T^\dagger x)_H = (x, \overline{\lambda} x)_H = \lambda$ so that $\lambda \in \mathrm{Num}(T)$ and the proof is finished. □

The following theorem due to F. Riesz establishes the properties of the spectral projection. The following version, which we state without proof, is due to Davies [14, Theorem 1.5.4, Theorem 1.5.1].

Theorem A.8 (Spectral projection). *Let V be a Banach space and T be a bounded linear operator on V. Let $\gamma \subset \mathbb{C}$ be a closed Jordan curve, i.e., γ is an image of a continuous map $\varphi : [0,1] \to \mathbb{C}$ such that $\varphi|_{[0,1)}$ is injective and $\varphi(0) = \varphi(1)$. Suppose that γ encloses a compact component S of $\sigma(T)$ and $\sigma(T) \setminus S$ is outside of γ. Then*

$$P_S \coloneqq -\frac{1}{2\pi i} \int_\gamma (T - zI)^{-1} \,\mathrm{d}z$$

is a bounded projection on V, which commutes with T. It is independent of the choice of γ subject to the aforementioned conditions. The restriction of T to $P_S V$ has spectrum S and the restriction of T to $(I - P_S)V$ has spectrum $\sigma(T) \setminus S$.

Now we proceed by presenting the spectral theory of the Fredholm operators, largely following Davies [14, section 4.3].

Definition A.9 (Fredholm operators, essential spectrum). *Let V, W be Banach spaces. A bounded linear operator $T : V \to W$ is said to be a* Fredholm operator *if its kernel and its cokernel,*

$$\begin{aligned} \operatorname{Ker}(T) &= \{x \in V,\, Tx = 0\}, \\ \operatorname{Coker}(T) &\coloneqq W/TV = \{\{y + z,\, z \in TV\},\, y \in W\}, \end{aligned}$$

are both finite-dimensional. The index of T is defined by

$$\operatorname{Ind}(T) \coloneqq \dim \operatorname{Ker}(T) - \dim \operatorname{Coker}(T).$$

The essential spectrum of $T \in \mathcal{L}(V)$ is defined as

$$\sigma_{\mathrm{ess}}(T) \coloneqq \{\lambda \in \mathbb{C},\, T - \lambda I \textit{ is not Fredholm}\}.$$

In the following theorem we collect some facts about essential spectra that can be found, including proofs, in [14, Corollary 4.3.8, Theorem 4.3.18].

Theorem A.10. *Let V be a Banach space and T be a bounded linear operator on V. Then for any compact linear operator K on V it holds true that*

$$\sigma_{\mathrm{ess}}(T) = \sigma_{\mathrm{ess}}(T + K).$$

Denote the unbounded component of $\mathbb{C} \setminus \sigma_{\mathrm{ess}}(T)$ by $\mathbb{U}$. Then, for every $\lambda \in \mathbb{U}$, $T - \lambda I$ is a Fredholm operator and $\operatorname{Ind}(T - \lambda I) = 0$. *Furthermore, $\sigma(T) \cap \mathbb{U}$ consists of at most countably many eigenvalues of finite algebraic and geometric multiplicities; accumulation points of $\sigma(T) \cap \mathbb{U}$, if any, are located only on $\partial\sigma_{\mathrm{ess}}$.*

This immediately implies the spectral theory of compact operators first developed by F. Riesz, i.e., that if V is an infinite-dimensional Banach space and $K \in \mathcal{L}(V)$ is compact, then $\sigma_{\mathrm{ess}}(K) = \{0\}$ and $\sigma(K) \setminus \{0\}$ consists of at most countably many eigenvalues of finite multiplicity, which can only accumulate at $\{0\}$.

In the rest of the section we present recent results which quantify the rate of accumulation for a special class of operators, namely bounded linear operators $A + K$ with A self-adjoint and K compact but not necessarily self-adjoint. First we need a measure of compactness, so we introduce approximation numbers of linear operators and the space of p-Schatten operators. Approximation numbers of a bounded linear operator $T : V_1 \to V_2$ between Banach spaces V_1, V_2 are defined by

$$a_j(T) = \inf_{\substack{M \in \mathcal{L}(V_1, V_2) \\ \operatorname{rank}(M) < j}} \|T - M\|_{\mathcal{L}(V_1, V_2)}, \qquad j = 1, 2, \dots \tag{A.9}$$

The approximation numbers fulfill, see Pietsch [53, paragraph 2.2.1, p. 79, Theorem 2.3.3, p. 83],

$$\|T\|_{\mathcal{L}(V_1,V_2)} = a_1(T) \geq a_2(T) \geq \ldots \geq 0, \tag{A.10a}$$

$$a_{j+k-1}(S+T) \leq a_j(S) + a_k(T), \qquad j,k = 1,2,\ldots, \tag{A.10b}$$

$$a_j(WTU) \leq \|W\|_{\mathcal{L}(V_2,V_3)} a_j(T) \|U\|_{\mathcal{L}(V_0,V_1)}, \qquad j = 1,2,\ldots \tag{A.10c}$$

with any $U \in \mathcal{L}(V_0,V_1)$, $S,T \in \mathcal{L}(V_1,V_2)$, $W \in \mathcal{L}(V_2,V_3)$ on Banach spaces $V_0,\ldots,V_3$.

Definition A.11 (p-Schatten class)**.** *A linear operator $T : V_1 \to V_2$ is said to be of p-Schatten class for some $1 \leq p < \infty$, symbolically $T \in \mathcal{S}_p(V_1,V_2)$, if its p-Schatten norm*

$$\|T\|_{\mathcal{S}_p(V_1,V_2)} = \Big(\sum_{j=1}^{\infty} a_j(T)^p\Big)^{\frac{1}{p}} \tag{A.11}$$

is finite.

Class $\mathcal{S}_p(V_1,V_2)$ is a subspace of compact operators from V_1 to V_2, a Banach space with respect to norm (A.11), and an operator ideal; see [53]. Notation $\mathcal{S}_p$ can be used instead of $\mathcal{S}_p(V_1,V_2)$ for brevity if the choice of spaces is clear.

We continue with two recent results which can be seen as a generalization of Kato's result [32] to perturbations which are not necessarily self-adjoint.

Theorem A.12 (Hansmann [28, Theorem 2.1])**.** *Let A, B be bounded operators on a Hilbert space H, let A be self-adjoint, and $B - A \in \mathcal{S}_p(H)$ for some $p \geq 1$. Then*

$$\sum_{\lambda \in \sigma_{\mathrm{p}}(B)} \operatorname{dist}\big(\lambda, \overline{\operatorname{Num}}(A)\big)^p \leq \|B-A\|^p_{\mathcal{S}_p(H)},$$

where each eigenvalue is counted according to its algebraic multiplicity.

We present the proof from [28]:

Proof. Note that the Schatten norm (A.11) can be characterized, for $p \geq 1$, by

$$\|K\|^p_{\mathcal{S}_p(H)} = \sup_{\{f_j\}_{j=1}^{\infty},\, \{g_j\}_{j=1}^{\infty}} \sum_{j=1}^{\infty} |(Kf_j, g_j)|^p, \tag{A.12}$$

where the supremum is taken over all extended orthonormal sequences (sequences which are either orthonormal or finite, extended by zeros) in H; see [53, Lemma 2.11.12, Proposition on p. 127].

Another needed tool is Schur's lemma, which roughly says that that the spectral projection of a bounded linear operator corresponding to (some of) its isolated eigenvalues with finite multiplicity is unitarily similar to a triangular operator in ℓ^2. Precisely, for $T \in \mathcal{L}(H)$ there exists a double sequence $\{t_{jk}\}_{j,k=1}^{\infty} \subset \mathbb{C}$ with $t_{jk} = 0$ for $j > k$, $t_{jj} = \lambda_j$ and an extended orthonormal sequence $\{f_j\}_{j=1}^{\infty} \subset H$ such that

$$Tf_j = t_{j1}f_1 + t_{j2}f_2 + \cdots + t_{jj}f_j \qquad \text{for all } j = 1,2,\ldots, \tag{A.13}$$

where $\{\lambda_j\}_{j=1}^{\infty}$ is a subset of eigenvalues of T which have finite algebraic multiplicity; see [25, Remark 4.1]. After all, this result is easily deduced from Theorem A.8 and Schur's decomposition for matrices.

With λ_j, f_j of (A.13) applied to $T := B$ we have, using (A.12) with $K := B - A$,

$$\|B-A\|^p_{\mathcal{S}_p(H)} \geq \sum_{j=1}^{\infty} |((B-A)f_j, f_j)|^p = \sum_{j=1}^{\infty} |\lambda_j - (Af_j, f_j)|^p$$

$$\geq \sum_{j=1}^{\infty} \operatorname{dist}(\lambda_j, \operatorname{Num}(A))^p,$$

where the sum is over all eigenvalues of B with finite algebraic multiplicity while repeating the eigenvalues according to their algebraic multiplicity. By virtue of Theorem A.10, the eigenvalues of infinite multiplicity are only located in $\overline{\mathrm{Num}}(A)$ and hence the sum can be extended to all eigenvalues of B and the proof is finished. □

Theorem A.13 (Hansmann [29, Corollary 1])**.** *Let A, B be bounded operators on a Hilbert space H, let A be self-adjoint, and $B - A \in \mathcal{S}_p(H)$ for some $p > 1$. Then*

$$\sum_{\lambda\in\sigma_{\mathrm{p}}(B)} \operatorname{dist}\big(\lambda, \sigma(A)\big)^p \leq C_p \|B - A\|^p_{\mathcal{S}_p(H)},$$

where each eigenvalue is counted according to its algebraic multiplicity. The constant $C_p \geq 2$ depends only on p; in particular it is independent of H.

Notice that Theorem A.12 does not feature the multiplicative constant $C_p \geq 2$ and its proof is very simple. Observe that, in particular when $p = 6$, which is our application in Theorem 2.17, the proof of Theorem A.13 in [29] uses C_6 which is known to fulfill

$$7.05 \times 10^6 \leq C_6 \leq 1.71 \times 10^7.$$

On the other hand, Theorem A.13 gives finer infomation about the accumulation around non-convex $\sigma(A)$.

Appendix B Stability of contractive GMRES convergence under compact perturbations

The purpose of this section is to extend the result of Moret [48] to show that when operators which are subject to contractive GMRES convergence[9] are compactly perturbed, contractive GMRES convergence with the same contraction factor is preserved up to a superlinearly quickly vanishing delay. A measure of compactness determines the convergence rate of the delay. Moret's result [48], as a special case, considers only compact perturbations of the identity; in this case the convergence rate for the perturbed operator is superlinear; see Remark B.5.

Consider a complex Banach space V and operator $T \in \mathcal{L}(V)$. Let $u_0, b \in V$ such that b is in the range of T. Denote $r_0 \coloneqq b - Tu_0$. The GMRES (*generalized minimal residual*) algorithm constructs a sequence $\{u_k\}_{k=1}^\infty \subset V$ given by

$$u_k = u_0 + \hat{p}_k(T) r_0, \tag{B.1a}$$
$$\hat{p}_k = \operatorname*{arg\,min}_{\hat{p}\in\mathcal{P}_{k-1}} \|r_0 - T\hat{p}(T) r_0\|_V, \tag{B.1b}$$

where $\mathcal{P}_k$ is the space of polynomials of degree at most k with complex coefficients. It is easy to see that if T is invertible and V is strictly convex, relations (B.1) in fact fully and uniquely determine $\{u_k\}_{k=1}^\infty$. It is convenient for notation to rewrite (B.1b) as

$$p_k = \operatorname*{arg\,min}_{p\in\mathcal{P}_k,\, p(0)=1} \|p(T) r_0\|_V. \tag{B.2}$$

The polynomials $\hat{p}_k$ are then recovered by $\hat{p}_k(t) = \frac{1-p_k(t)}{t}$. Denoting $r_k \coloneqq b - Tu_k$ yields

$$\|r_k\|_V = \min_{p_k\in\mathcal{P}_k,\, p_k(0)=1} \|p_k(T) r_0\|_V \leq \|r_0\|_V \min_{p_k\in\mathcal{P}_k,\, p_k(0)=1} \|p_k(T)\|_{\mathcal{L}(V)}. \tag{B.3}$$

The equality in (B.3) motivates the name of the method: r_k are the residuals to be minimized.

[9] We say that operator T on a Banach space V is subject to *contractive GMRES convergence* with a contraction factor $M \in [0, 1)$ if for every initial residual $r_0 \in V$ there is a decrease in GMRES residual norm (B.3) by factor M in every step, i.e.,

$$\frac{\|r_k\|_V}{\|r_{k-1}\|_V} \leq M \qquad \text{for all } k \in \mathbb{N} \text{ and every initial residual } r_0 \in V.$$

In the remainder of this section we assume that V is a Hilbert space with an inner product $(\cdot,\cdot)$, T is a bounded linear operator on V, and the symbol $\|\cdot\|$ stands for either the norm on V induced by $(\cdot,\cdot)$ or the induced operator norm on $\mathcal{L}(V)$. For $k=0,1,2,\dots$ define a Krylov subspace $\mathcal{K}_k := \operatorname{span}\{r_0, Tr_0, T^2r_0, \dots, T^{k-1}r_0\} \subset V$. Then the GMRES algorithm characterized by (B.1) can be equivalently described as

$$u_k = \operatorname*{arg\,min}_{u_k\in V;\, u_k-u_0\in\mathcal{K}_k} \|f - Tu_k\|. \tag{B.4}$$

Assume that $t_1, t_2, \dots, t_k$ is the orthonormal basis of $\mathcal{K}_k$, $k=1,2,\dots$, and $z_1, z_2, \dots, z_k$ is the orthornormal basis of $T\mathcal{K}_k$, $k=1,2,\dots$. This is well-defined if

$$\mathcal{K}_{k+1} \supsetneq \mathcal{K}_k \qquad \text{for all } k=1,2,\dots. \tag{B.5}$$

It is well-known that in the converse case, when $\mathcal{K}_{m+1} = \mathcal{K}_m \supsetneq \mathcal{K}_{m-1} \supsetneq \dots$ for certain m, the solution has been reached, i.e., $Tu_m = b$, provided that T is invertible. To see this, observe that $T\mathcal{K}_m \subset \mathcal{K}_{m+1}$ but at the same time $\dim T\mathcal{K}_m = \dim\mathcal{K}_m = \dim\mathcal{K}_{m+1}$ by the invertibility of T; hence $T\mathcal{K}_m = \mathcal{K}_{m+1} \ni r_0$, i.e., $r_0 = \hat{p}(T)Tr_0$ with some $\hat{p}\in\mathcal{P}_{m-1}$ which means that $r_m = r_0 - \hat{p}(T)Tr_0 = 0$. All following convergence results will cover this situation as a special case. Hence we can assume from now on, without loss of generality, that (B.5) holds.

Moret [48] proves the following auxiliary result.

Lemma B.1 (Moret [48, Lemma 6])**.** *Let T be invertible. For every $k\in\mathbb{N}$ and $\lambda\in\mathbb{C}$ it holds*

$$\|r_k\| = |(t_{k+1}, z_k)|\|r_{k-1}\| = |(t_{k+1}, (I-\lambda T^{-1})z_k)|\|r_{k-1}\|. \tag{B.6}$$

Note that the second equality of (B.6) follows trivially by construction.

Lemma B.2 (Pietsch [53, Lemma 2.11.13, p. 125])**.** *Let T be a bounded linear operator on a Hilbert space H. Let $a_1(T) \ge a_2(T) \ge a_3(T) \ge \dots \ge 0$ denote the approximation numbers of T as defined by* (A.9)*. Then for any pair of orthonormal families $\{f_1, f_2, \dots, f_k\}$, $\{g_1, g_2, \dots, g_k\} \subset H$ it holds that*

$$\det\big\{(Tf_i, g_j)\big\}_{i,j=1}^k \le \prod_{j=1}^k a_k(T). \tag{B.7}$$

Now we provide a modification of [48, equation (2.7)].

Lemma B.3. *Let T be invertible and $\lambda\in\mathbb{C}$ be arbitrary. Denote the approximation numbers of $I-\lambda T^{-1}$ by $a_1(I-\lambda T^{-1}) \ge a_2(I-\lambda T^{-1}) \ge \dots \ge 0$. Then*

$$\tfrac{\|r_k\|}{\|r_0\|} \le \prod_{j=1}^k a_j(I-\lambda T^{-1}). \tag{B.8}$$

Proof. From (B.6) we have

$$\tfrac{\|r_k\|}{\|r_0\|} = \prod_{j=1}^k |(t_{j+1}, (I-\lambda T^{-1})z_j)|.$$

The matrix

$$\big\{|(t_{i+1}, (I-\lambda T^{-1})z_j)|\big\}_{i,j=1}^k$$

is upper triangular because by construction,

$$\begin{aligned} 0 &= (t_{j+2}, z_j) = (t_{j+3}, z_j) = \dots, \\ 0 &= (t_{j+1}, T^{-1}z_j) = (t_{j+2}, T^{-1}z_j) = \dots. \end{aligned}$$

This implies that $\prod_{j=1}^k |(t_{j+1}, (I-\lambda T^{-1})z_j)| = \det\big\{|(t_{i+1}, (I-\lambda T^{-1})z_j)|\big\}_{i,j=1}^k$ which is bounded by $\prod_{j=1}^k a_j(I-\lambda T^{-1})$ due to (B.7). □

Now we are in the position to characterize GMRES convergence for a (non-self-adjoint) compact perturbation of a self-adjoint operator.

Theorem B.4. *Let $T = B + C$ be an invertible linear operator on a Hilbert space. Let B be self-adjoint and positive, with spectrum $\sigma(B) \subset [a, b] \subset (0, \infty)$, and let C be compact. Then GMRES iterations with T and r_0 produce residuals r_k with the norm*

$$\|r_k\| = \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(T)r_0\|$$

which fulfills

$$\limsup_{k\to\infty} \tfrac{\|r_k\|}{\|r_{k-1}\|} \leq \tfrac{b-a}{b+a}. \tag{B.9}$$

Furthermore, if C is of p-Schatten class for some $p \geq 1$, then there exists $c \geq 0$ independent of r_0 such that

$$\big(\tfrac{\|r_k\|}{\|r_0\|}\big)^{\frac{1}{k}} \leq \tfrac{b-a}{b+a} + ck^{-\frac{1}{p}}. \tag{B.10}$$

The constant c fulfills the bound

$$c \leq \tfrac{2b}{b+a}\|T^{-1}\|\|C\|_{\mathcal{S}_p}. \tag{B.11}$$

Proof. Set $\lambda \coloneqq \frac{2}{a^{-1}+b^{-1}}$. For the spectrum of $I-\lambda B^{-1}$ we then have $\sigma(I-\lambda B^{-1}) \subset [-\frac{b-a}{b+a}, \frac{b-a}{b+a}]$. Thanks to the assumptions we can express $T^{-1} = B^{-1} - B^{-1}CT^{-1}$. Hence by Lemma B.1,

$$\begin{aligned}\tfrac{\|r_k\|}{\|r_{k-1}\|} &\leq |(t_{k+1}, (I - \lambda B^{-1})z_k)| + |(t_{k+1}, \lambda B^{-1}CT^{-1}z_k)|\\ &\leq \tfrac{b-a}{b+a} + \tfrac{\lambda}{a}\|CT^{-1}z_k\|.\end{aligned}$$

The sequence $\{z_k\}_{k=1}^{\infty}$ is an orthonormal system which is, by Bessel's inequality, weakly null. By the compactness of C, (B.9) follows.

Denote $M \coloneqq \frac{b-a}{b+a} \geq \|I - \lambda B^{-1}\|$. By (B.8), the AM-GM inequality, and the Minkowski inequality we obtain

$$\big(\tfrac{\|r_k\|}{\|r_0\|}\big)^{\frac{1}{k}} \leq \prod_{j=1}^{k} a_j(I - \lambda T^{-1})^{\frac{1}{k}} \leq \tfrac{1}{k}\sum_{j=1}^{k} a_j(I - \lambda T^{-1}) \tag{B.12}$$

$$= M + \tfrac{1}{k}\sum_{j=1}^{k}\big(a_j(I - \lambda T^{-1}) - M\big)$$

$$\leq M + k^{-\frac{1}{p}}\left(\sum_{j=1}^{k}\big(a_j(I - \lambda T^{-1}) - M\big)^p\right)^{\frac{1}{p}}. \tag{B.13}$$

By the properties of approximation numbers (A.10) and $\lambda\|B^{-1}\| \leq (1 + M)$ we get

$$\begin{aligned}a_j(I - \lambda T^{-1}) &= a_j(I - \lambda B^{-1} + \lambda B^{-1}CT^{-1})\\ &\leq a_1(I - \lambda B^{-1}) + a_j(\lambda B^{-1}CT^{-1}) \leq M + (1 + M)\|T^{-1}\|a_j(C).\end{aligned}$$

Henceforth, with the aid of (A.11),

$$\begin{aligned}\sum_{j=1}^{k}\big(a_j(I - \lambda T^{-1}) - M\big)^p &\leq (1 + M)^p\|T^{-1}\|^p\sum_{j=1}^{k} a_j(C)^p\\ &\leq (1 + M)^p\|T^{-1}\|^p\|C\|_{\mathcal{S}_p}^p,\end{aligned}$$

which shows that the sum in (B.13) is bounded independently of k and the estimate (B.10) with (B.11) follows. □

Remark B.5. *Notice that the result of Moret [48] for $B = \lambda I$, $\lambda \neq 0$, follows as a special case of Theorem B.4 by setting $a = b = \lambda$;* (B.9) *gives the* q-superlinear convergence $\lim_{k\to\infty} \frac{\|r_k\|}{\|r_{k-1}\|} = 0$ *of [48, Theorem 1] and* (B.10) *gives the rate $\|r_k\|^{\frac{1}{k}} = O(k^{-\frac{1}{p}})$ of [48, equation (1.1)].*

Remark B.6. *Theorem B.4 holds for non-self-adjoint B provided*

$$M := \|I - \lambda B^{-1}\| < 1 \qquad \textit{for some } \lambda \in \mathbb{C}. \tag{B.14}$$

Then (B.9)–(B.11) *are replaced by*

$$\limsup_{k\to\infty} \tfrac{\|r_k\|}{\|r_{k-1}\|} \le M, \qquad \big(\tfrac{\|r_k\|}{\|r_0\|}\big)^{\frac{1}{k}} \le M + ck^{-\frac{1}{p}}, \qquad c \le (1+M)\|T^{-1}\|\|C\|_{\mathcal{S}_p}, \tag{B.15}$$

respectively. We leave the obvious modification of the proof to the reader. If B is normal, a sufficient and necessary condition for validity of (B.14) *is*

$$\overline{\mathrm{Num}} B^{-1} \subset \mathcal{B}_{M|\lambda^{-1}|}(\lambda^{-1}) \qquad \textit{for some } \lambda \in \mathbb{C} \textit{ and } M < 1,$$

where $\mathcal{B}_R(z)$ is a closed ball of diameter R and center z in the complex plane. For a non-normal operator B it holds, see Horn and Johnson [31, Problem 5.7.P20], that

$$\|I - \lambda B^{-1}\| \le 2 \sup_{z \in \overline{\mathrm{Num}}(I - \lambda B^{-1})} |z|,$$

and hence a sufficient condition for the validity of (B.14) *for non-normal B is*

$$\overline{\mathrm{Num}} B^{-1} \subset \mathcal{B}_{\frac{M}{2}|\lambda^{-1}|}(\lambda^{-1}) \qquad \textit{for some } \lambda \in \mathbb{C} \textit{ and } M < 1.$$

Remark B.7. *Nevanlinna [50, Theorem 1.2] shows in a setting similar to* (B.14)*, when B and C are not necessarily self-adjoint and C is additionally of* 1*-Schatten class, that for any $\epsilon > 0$ there exist $C_\epsilon > 0$ and $k_0 \in \mathbb{N}$ such that*

$$\big(\tfrac{\|r_k\|}{C_\epsilon \|r_0\|}\big)^{\frac{1}{k}} \le M + \epsilon \qquad \textit{for all } k \ge k_0$$

and $C_\epsilon \to \infty$ with $\epsilon \to 0+$. It is not difficult to see that this result is covered by (B.15)*. Malinen [46, Lemma 6.8] similarly deals with polynomial iterations of $B + C$ with C of p-Schatten class, $p > 1$, and provides only a result, which is valid only asymptotically, for large enough k. Both of these results use methods of complex analysis to study iterations using* monic *polynomials and subsequent normalization to $p(0) = 1$ gives only asymptotic behavior.*

On the other hand, Nevanlinna [50, Theorem 4.2] provides a result valid for all $k \in \mathbb{N}$:

$$\big(\tfrac{\|r_k\|}{C_\eta \|r_0\|}\big)^{\frac{1}{k}} \le \eta$$

where η is determined in terms of the capacity of $\sigma(B)$; the disadvantage of this estimate is that the $ck^{-\frac{1}{p}}$ term of (B.15) *is forgotten at the cost of a potentially large constant C_η, which does not pollute* (B.15)*.*

Remark B.8. *Herzog and Sachs [30, Theorem 3.12] provide an estimate similar to* (B.12) *in the context of MINRES for self-adjoint T such that $T - \lambda I$ is compact for some $\lambda \neq 0$ (or, more generally, $p(T)T - I$ is compact for some polynomial p); specifically,*

$$\big(\tfrac{\|r_k\|}{\|r_0\|}\big)^{\frac{1}{k}} \le \tfrac{2c}{k} \sum_{j=1}^{k} |\lambda_j(C)|$$

where $\lambda_j(C)$ are distinct *eigenvalues of C ordered by decreasing magnitude. A notable property is that the eigenvalues are not taken according to multiplicity, i.e., repeated eigenvalues contribute just once, which seems to be only possible under normality of T.*

Now let us consider how to prevent the influence of isolated eigenvalues of B on the contraction factor $\frac{b-a}{b+a}$ in Theorem B.4. Assume $S \subset \sigma_{\mathrm{p}}(T)$ is a set of finitely many isolated eigenvalues of finite multiplicity and define P_S to be the spectral projection provided by Theorem A.8. Then we can write $T = B + C = B(I - P_S) + BP_S + C$. The projector P_S has finite rank, and hence $BP_S + C$ is compact. The operator $B(I - P_S)$ might have a better spectral bound than B so one could expect a better contraction factor in (B.9) and (B.10). But we cannot apply Theorem (B.4) directly as $B(I - P_S)$ is singular unless $S = \emptyset$. The following theorem gives a composite bound which removes the contribution of the isolated eigenvalues to the contraction factor.

Theorem B.9. *Let $T = B + C$ be an invertible linear operator on a Hilbert space. Let B be a bounded linear operator with spectrum $\sigma(B) \subset S \cup [\hat{a}, \hat{b}]$ with $0 < \hat{a} \leq \hat{b} < \infty$ and $S \subset \sigma_{\mathrm{p}}(B) \subset \mathbb{C} \setminus \{0\}$ consisting of only a finite number of isolated eigenvalues of finite multiplicity. Assume that C is compact. Then GMRES iterations with T and r_0 produce residuals r_k with norms*

$$\|r_k\| = \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|p(T)r_0\|,$$

which fulfill

$$\limsup_{k\to\infty} \tfrac{\|r_k\|}{\|r_{k-1}\|} \leq \tfrac{\hat{b}-\hat{a}}{\hat{b}+\hat{a}}. \tag{B.16}$$

Furthermore, if C is of p-Schatten class for some $p \geq 1$, then there exists $\hat{c} \geq 0$ independent of r_0 such that

$$\big(\tfrac{\|r_k\|}{\|r_0\|}\big)^{\frac{1}{k}} \leq \tfrac{\hat{b}-\hat{a}}{\hat{b}+\hat{a}} + \hat{c}k^{-\frac{1}{p}}. \tag{B.17}$$

Define P_S as the spectral projection corresponding to S as defined by Theorem A.8. Set $\hat{\lambda} := \frac{2}{\hat{a}^{-1}+\hat{b}^{-1}}$. Then the constant $\hat{c}$ fulfills the bound

$$\hat{c} \leq \|\hat{\lambda}B^{-1}CT^{-1} + (I - \hat{\lambda}B^{-1})P_S\|_{\mathcal{S}_p} < \infty. \tag{B.18}$$

Proof. Thanks to the assumptions we can express $T^{-1} = B^{-1} - B^{-1}CT^{-1}$. With $\hat{\lambda}$ and P_S from the statement of the theorem we have

$$I - \hat{\lambda}T^{-1} = (I - \hat{\lambda}B^{-1})(I - P_S) + \hat{\lambda}B^{-1}CT^{-1} + (I - \hat{\lambda}B^{-1})P_S. \tag{B.19}$$

By virtue of Theorem A.8 we have $\sigma((I - \hat{\lambda}B^{-1})(I - P_S)) \subset [-\frac{\hat{b}-\hat{a}}{\hat{b}+\hat{a}}, \frac{\hat{b}-\hat{a}}{\hat{b}+\hat{a}}]$ and hence $\|(I - \hat{\lambda}B^{-1})(I - P_S)\| \leq \frac{\hat{b}-\hat{a}}{\hat{b}+\hat{a}} =: \hat{M}$. By the assumptions, C and P_S are compact and hence by the same arguments as in the proof of Theorem B.4 we obtain (B.16).

By the properties of approximation numbers (A.10) and using (B.19) we obtain

$$a_j(I - \hat{\lambda}T^{-1}) \leq \underbrace{\|(I - \hat{\lambda}B^{-1})(I - P_S)\|}_{\leq \hat{M}} + a_j(\hat{\lambda}B^{-1}CT^{-1} + (I - \hat{\lambda}B^{-1})P_S)$$

so that, with the aid of (A.11), we get

$$\begin{aligned}\sum_{j=1}^{k}\big(a_j(I - \hat{\lambda}T^{-1}) - \hat{M}\big)^p &\leq \sum_{j=1}^{k} a_j(\hat{\lambda}B^{-1}CT^{-1} + (I - \hat{\lambda}B^{-1})P_S)^p \\ &\leq \|\hat{\lambda}B^{-1}CT^{-1} + (I - \hat{\lambda}B^{-1})P_S\|_{\mathcal{S}_p}^p,\end{aligned}$$

where the right-hand side of the last inequality is finite because $C \in \mathcal{S}_p$ and P_S has finite rank. Using (B.13), but with $\hat{M}$, $\hat{\lambda}$ instead of M, λ, respectively, we obtain (B.17), (B.18) and the proof is finished. □

Remark B.10. *Analogously to Theorem B.9, which improves Theorem B.4 by removing the contribution of a finite number of isolated eigenvalues to the contraction factor, for B which is not self-adjoint one can improve the bound of Remark B.6 to*

$$\begin{gathered}\limsup_{k\to\infty} \tfrac{\|r_k\|}{\|r_{k-1}\|} \le \hat{M}, \qquad \big(\tfrac{\|r_k\|}{\|r_0\|}\big)^{\frac{1}{k}} \le \hat{M} + \hat{c}k^{-\frac{1}{p}}, \\ \hat{c} \le \|\hat{\lambda}B^{-1}CT^{-1} + (I - \hat{\lambda}B^{-1})P_S\|_{\mathcal{S}_p},\end{gathered} \tag{B.20}$$

which is valid provided that, instead of (B.14)*, it holds*

$$\hat{M} := \|(I - \hat{\lambda}B^{-1})(I - P_S)\| < 1 \quad \textit{for some } \hat{\lambda} \in \mathbb{C}. \tag{B.21}$$

The proof involves a minimal modification of the proof of Theorem B.9 and we leave it as an exercise.

Remark B.11. *It seems plausible that Theorem B.9 would hold for countable sets S provided that S accumulates at $[\hat{a}, \hat{b}]$. Proving this might require a finer, or different, approach.*

References

[1] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. Rognes, and G. Wells. “The FEniCS Project Version 1.5”. In: *Archive of Numerical Software* 3.100 (2015). DOI: 10.11588/ans.2015.100.20553.

[2] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, V. Eijkhout, W. Gropp, D. Kaushik, M. Knepley, D. May, L. C. McInnes, R. T. Mills, T. Munson, K. Rupp, P. Sanan, B. Smith, S. Zampini, H. Zhang, and H. Zhang. *PETSc Users Manual.* Tech. rep. ANL-95/11 - Revision 3.10. Argonne National Laboratory, 2018. URL: http://www.mcs.anl.gov/petsc.

[3] S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith. “Efficient Management of Parallelism in Object Oriented Numerical Software Libraries”. In: *Modern Software Tools in Scientific Computing.* Ed. by E. Arge, A. M. Bruaset, and H. P. Langtangen. Birkhäuser Press, 1997, pp. 163–202. URL: http://dx.doi.org/10.1007/978-1-4612-1986-6_8.

[4] Y. Berchenko-Kogan. “The View from Here: What Do Grad Students in Math Do All Day?” In: *Math Horizons* 20.3 (2013), pp. 18–19. DOI: 10.4169/mathhorizons.20.3.18. eprint: https://www.quora.com/Mathematics/What-do-grad-students-in-math-do-all-day/answer/Yasha-Berchenko-Kogan.

[5] J. Blechta, J. Málek, and M. Vohralík. “Localization of the $W^{-1,q}$ norm for local a posteriori efficiency”. In: *IMA J. Numer. Anal.* (Mar. 2019). DOI: 10.1093/imanum/drz002.

[6] J. Blechta and M. Řehoř. *FENaPack 2018.1.0.* July 2018. DOI: 10.5281/zenodo.1308015.

[7] M. Braack and P. B. Mucha. “Directional do-nothing condition for the Navier-Stokes equations”. In: *J. Comput. Math.* 32.5 (2014), pp. 507–521. DOI: 10.4208/jcm.1405-m4347.

[8] J. Cahouet and J.-P. Chabard. “Some fast 3D finite element solvers for the generalized Stokes problem”. In: *Internat. J. Numer. Methods Fluids* 8.8 (1988), pp. 869–895. DOI: 10.1002/fld.1650080802.

[9] P. Ciarlet Jr. and M. Vohralík. “Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients”. In: *ESAIM Math. Model. Numer. Anal.* 52.5 (2018), pp. 2037–2064. DOI: 10.1051/m2an/2018034.

[10] E. Cosserat and F. Cosserat. “Sur les équations de la théorie de l’élasticité”. In: *C. R. Acad. Sci. (Paris)* 126 (1898), pp. 1089–1091.

[11] M. Costabel. “On the limit Sobolev regularity for Dirichlet and Neumann problems on Lipschitz domains”. In: *arXiv preprint arXiv:1711.07179* (2017). URL: https://arxiv.org/abs/1711.07179.

[12] M. Costabel, M. Crouzeix, M. Dauge, and Y. Lafranche. “The inf-sup constant for the divergence on corner domains”. In: *Numerical Methods for Partial Differential Equations* 31.2 (2015), pp. 439–458. DOI: 10.1002/num.21916.

[13] L. D. Dalcin, R. R. Paz, P. A. Kler, and A. Cosimo. “Parallel distributed computing using Python”. In: *Advances in Water Resources* 34.9 (2011). New Computational Methods and Software Tools, pp. 1124–1139. DOI: 10.1016/j.advwatres.2011.04.013.

[14] E. B. Davies. *Linear Operators and their Spectra*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2007. DOI: 10.1017/CBO9780511618864.

[15] P. Deuring. “Eigenvalue Bounds for the Schur Complement with a Pressure Convection-diffusion Preconditioner in Incompressible Flow Computations”. In: *J. Comput. Appl. Math.* 228.1 (June 2009), pp. 444–457. DOI: 10.1016/j.cam.2008.10.017.

[16] L. Diening, C. Kreuzer, and E. Süli. “Finite element approximation of steady flows of incompressible fluids with implicit power-law-like rheology”. In: *SIAM J. Numer. Anal.* 51.2 (2013), pp. 984–1015. DOI: 10.1137/120873133.

[17] N. Dunford and J. T. Schwartz. *Linear operators. Part I.* Wiley Classics Library. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1988, pp. xiv+858.

[18] H. Elman, D. Loghin, and A. Wathen. “Preconditioning Techniques for Newton’s Method for the Incompressible Navier-Stokes Equations”. English. In: *BIT Numerical Mathematics* 43.5 (2003), pp. 961–974. DOI: 10.1023/B:BITN.0000014565.86918.df.

[19] H. C. Elman. “Preconditioning for the steady-state Navier-Stokes equations with low viscosity”. In: *SIAM J. Sci. Comput.* 20.4 (1999), pp. 1299–1316. DOI: 10.1137/S1064827596312547.

[20] H. C. Elman, D. J. Silvester, and A. J. Wathen. “Performance and analysis of saddle point preconditioners for the discrete steady-state Navier-Stokes equations”. In: *Numerische Mathematik* 90.4 (2002), pp. 665–688. DOI: 10.1007/s002110100300.

[21] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers : with Applications in Incompressible Fluid Dynamics: with Applications in Incompressible Fluid Dynamics*. 1st. Numerical Mathematics and Scientific Computation. Oxford University Press, 2005. URL: https://books.google.cz/books?id=N9KEgXMrMm8C.

[22] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. 2nd. Oxford University Press, 2014.

[23] H. C. Elman and R. S. Tuminaro. “Boundary Conditions in Approximate Commutator Preconditioners for the Navier-Stokes Equations”. In: *Electronic Transactions on Numerical Analysis* 35 (2009), pp. 257–280. URL: http://etna.mcs.kent.edu/volumes/2001-2010/vol35/abstract.php?vol=35&pages=257-280.

[24] V. Faber, J. Liesen, and P. Tichý. “Properties of worst-case GMRES”. In: *SIAM J. Matrix Anal. Appl.* 34.4 (2013), pp. 1500–1519. DOI: 10.1137/13091066X.

[25] I. C. Gohberg and M. G. Kreĭn. *Introduction to the theory of linear nonselfadjoint operators*. Translated from the Russian by A. Feinstein. Translations of Mathematical Monographs, Vol. 18. American Mathematical Society, Providence, R.I., 1969, pp. xv+378.

[26] A. Greenbaum, V. Pták, and Z. Strakoš. “Any Nonincreasing Convergence Curve is Possible for GMRES”. In: *SIAM Journal on Matrix Analysis and Applications* 17.3 (1996), pp. 465–469. DOI: 10.1137/S0895479894275030.

[27] K. Gustafson. “The Toeplitz-Hausdorff theorem for linear operators”. In: *Proc. Amer. Math. Soc.* 25 (1970), pp. 203–204. DOI: 10.2307/2036559.

[28] M. Hansmann. “An eigenvalue estimate and its application to non-selfadjoint Jacobi and Schrödinger operators”. In: *Lett. Math. Phys.* 98.1 (2011), pp. 79–95. DOI: 10.1007/s11005-011-0494-9.

[29] M. Hansmann. “Variation of discrete spectra for non-selfadjoint perturbations of selfadjoint operators”. In: *Integral Equations Operator Theory* 76.2 (2013), pp. 163–178. DOI: 10.1007/s00020-013-2057-1.

[30] R. Herzog and E. Sachs. “Superlinear convergence of Krylov subspace methods for self-adjoint problems in Hilbert space”. In: *SIAM J. Numer. Anal.* 53.3 (2015), pp. 1304–1324. DOI: 10.1137/140973050.

[31] R. A. Horn and C. R. Johnson. *Matrix analysis.* Second. Cambridge University Press, Cambridge, 2013, pp. xviii+643.

[32] T. Kato. “Variation of discrete spectra”. In: *Comm. Math. Phys.* 111.3 (1987), pp. 501–504. URL: http://projecteuclid.org/euclid.cmp/1104159643.

[33] D. Kay and D. Loghin. *A Green’s function preconditioner for the Navier-Stokes equations.* Tech. rep. NA-99/06. Oxford University Computing Laboratory, 1999. URL: http://eprints.maths.ox.ac.uk/1292/1/NA-99-06.pdf.

[34] D. Kay, D. Loghin, and A. Wathen. “A Preconditioner for the Steady-State Navier–Stokes Equations”. In: *SIAM Journal on Scientific Computing* 24.1 (2002), pp. 237–256. DOI: 10.1137/S106482759935808X.

[35] A. Klawonn and G. Starke. “Block triangular preconditioners for nonsymmetric saddle point problems: field-of-values analysis”. In: *Numer. Math.* 81.4 (1999), pp. 577–594. DOI: 10.1007/s002110050405.

[36] A. Kozhevnikov. “A history of the Cosserat spectrum”. In: *The Maz’ya Anniversary Collection.* Ed. by J. Rossmann, P. Takáč, and G. Wildenhain. Basel: Birkhäuser Basel, 1999, pp. 223–234. DOI: 10.1007/978-3-0348-8675-8_16.

[37] O. A. Ladyzhenskaya. *The mathematical theory of viscous incompressible flow.* Revised English edition. Translated from the Russian by Richard A. Silverman. Gordon and Breach Science Publishers, New York-London, 1963, pp. xiv+184.

[38] M. Lanzendörfer and J. Hron. “On multiple solutions to the steady flow of incompressible fluids subject to do-nothing or constant traction boundary conditions on artificial boundaries”. In: *arXiv preprint arXiv:1904.04898* (2019). URL: https://arxiv.org/abs/1904.04898.

[39] J. Liesen and P. Tichý. “Convergence analysis of Krylov subspace methods”. In: *GAMM Mitt. Ges. Angew. Math. Mech.* 27.2 (2004), 153–173 (2005). DOI: 10.1002/gamm.201490008.

[40] J. Liesen and P. Tichý. “The worst-case GMRES for normal matrices”. In: *BIT* 44.1 (2004), pp. 79–98. DOI: 10.1023/B:BITN.0000025083.59864.bd.

[41] A. Logg, K.-A. Mardal, and G. Wells. *Automated solution of differential equations by the finite element method: The FEniCS book.* Vol. 84. Free download http://launchpad.net/fenics-book/trunk/final/+download/fenics-book-2011-10-27-final.pdf. Springer Science & Business Media, 2012. DOI: 10.1007/978-3-642-23099-8.

[42] D. Loghin and A. J. Wathen. “Analysis of Preconditioners for Saddle-Point Problems”. In: *SIAM Journal on Scientific Computing* 25.6 (2004), pp. 2029–2049. DOI: 10.1137/S1064827502418203.

[43] D. Loghin. *Analysis of preconditioned Picard iterations for the Navier-Stokes equations.* Tech. rep. NA-01/10. Oxford University Computing Laboratory, 2001. URL: http://eprints.maths.ox.ac.uk/1241/1/NA-01-10.pdf.

[44] J. Málek, J. Nečas, M. Rokyta, and M. Růžička. *Weak and measure-valued solutions to evolutionary PDEs.* London: Chapman & Hall, 1996, pp. xii+317.

[45] J. Málek and Z. Strakoš. *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs.* SIAM Spotlight Series. Philadelphia: SIAM, Jan. 2015. URL: http://bookstore.siam.org/sl01/.

[46] J. Malinen. *On the properties for iteration of a compact operator with unstructured perturbation.* Tech. rep. A360. Helsinki University of Technology, Institute of Mathematics, 1996. URL: https://math.aalto.fi/~jmalinen/MyPSFilesInWeb/SmallCompact.pdf.

[47] I. Mitrea and M. Mitrea. "The Poisson Problem with Mixed Boundary Conditions in Sobolev and Besov Spaces in Non-Smooth Domains". English. In: *Transactions of the American Mathematical Society* 359.9 (2007), pp. 4143–4182. URL: http://www.jstor.org/stable/20161769.

[48] I. Moret. "A note on the superlinear convergence of GMRES". In: *SIAM J. Numer. Anal.* 34.2 (1997), pp. 513–516. DOI: 10.1137/S0036142993259792.

[49] M. F. Murphy, G. H. Golub, and A. J. Wathen. "A Note on Preconditioning for Indefinite Linear Systems". In: *SIAM Journal on Scientific Computing* 21.6 (2000), pp. 1969–1972. DOI: 10.1137/S1064827599355153.

[50] O. Nevanlinna. "Convergence of Krylov methods for sums of two operators". In: *BIT* 36.4 (1996), pp. 775–785. DOI: 10.1007/BF01733791.

[51] M. A. Olshanskii, J. Peters, and A. Reusken. "Uniform preconditioners for a parameter dependent saddle point problem with application to generalized Stokes interface equations". In: *Numer. Math.* 105.1 (2006), pp. 159–191. DOI: 10.1007/s00211-006-0031-4.

[52] M. A. Olshanskii and Y. V. Vassilevski. "Pressure Schur Complement Preconditioners for the Discrete Oseen Problem". In: *SIAM Journal on Scientific Computing* 29.6 (2007), pp. 2686–2704. DOI: 10.1137/070679776.

[53] A. Pietsch. *Eigenvalues and s-numbers.* Vol. 13. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1987, p. 360.

[54] R. Rannacher. "A short course on numerical simulation of viscous flow: discretization, optimization and stability analysis". In: *Discrete Contin. Dyn. Syst. Ser. S* 5.6 (2012), pp. 1147–1194. DOI: 10.3934/dcdss.2012.5.1147.

[55] M. Řehoř. "Diffuse interface models in theory of interacting continua". 2018. URL: http://hdl.handle.net/20.500.11956/103641.

[56] M. Řehoř, J. Blechta, and O. Souček. "On some practical issues concerning the implementation of Cahn–Hilliard–Navier–Stokes type models". In: *International Journal of Advances in Engineering Sciences and Applied Mathematics* (2016), pp. 1–10. DOI: 10.1007/s12572-016-0171-4.

[57] D. Silvester, H. Elman, D. Kay, and A. Wathen. "Efficient preconditioning of the linearized Navier–Stokes equations for incompressible flow". In: *Journal of Computational and Applied Mathematics* 128.1–2 (2001). Numerical Analysis 2000. Vol. VII: Partial Differential Equations, pp. 261–279. DOI: 10.1016/S0377-0427(00)00515-X.

[58] D. Silvester, H. Elman, and A. Ramage. *Incompressible Flow and Iterative Solver Software (IFISS), version 3.3.* Oct. 2013. URL: https://www.maths.manchester.ac.uk/~djs/ifiss/ifiss3.3.tar.gz.

[59] G. Stoyan. "Towards discrete Velte decompositions and narrow bounds for inf-sup constants". In: *Comput. Math. Appl.* 38.7-8 (1999), pp. 243–261. DOI: 10.1016/S0898-1221(99)00254-0.

[60] R. Temam. *Navier-Stokes equations.* Third. Vol. 2. Studies in Mathematics and its Applications. Theory and numerical analysis, With an appendix by F. Thomasset. North-Holland Publishing Co., Amsterdam, 1984, pp. xii+526.

[61] P. Tichý, J. Liesen, and V. Faber. "On worst-case GMRES, ideal GMRES, and the polynomial numerical hull of a Jordan block". In: *Electron. Trans. Numer. Anal.* 26 (2007), pp. 453–473. URL: http://etna.mcs.kent.edu/volumes/2001-2010/vol26/abstract.php?vol=26&pages=453-473.

[62] H. Triebel. *Theory of function spaces. III.* Vol. 100. Monographs in Mathematics. Birkhäuser Verlag, Basel, 2006, pp. xii+426.

[63] A. J. Wathen. "Realistic eigenvalue bounds for the Galerkin mass matrix". In: *IMA J. Numer. Anal.* 7.4 (1987), pp. 449–457. DOI: 10.1093/imanum/7.4.449.

[64] K. Yosida. *Functional analysis.* Classics in Mathematics. Reprint of the sixth (1980) edition. Springer-Verlag, Berlin, 1995, pp. xii+501. DOI: 10.1007/978-3-642-61859-8.

Conclusion

In Section I.2 we provided a novel view on the classification of incompressible fluids. Table I.1 displays the range of considered constitutive relations. A new class of activated Euler fluids, which includes the Euler/Navier-Stokes fluid (I.2.30) and the Euler/power-law fluid (I.2.31), has been considered; see also Figure I.6. In Section I.2.7 activated boundary conditions, e.g., stick-slip, have been investigated. A characterization of simple shear flows of the Euler/Navier-Stokes fluid is provided in Sections I.2.6 and I.2.7; see also Table I.3. In Section I.1 we discussed a potential employment of the Euler/Navier-Stokes fluid and other activated Euler fluids for the description of boundary layers. Section I.3 is devoted to large-data existence analysis of flows, both steady and unsteady, of activated Euler fluids under (i) the no-slip boundary condition and (ii) a variety of slip-like boundary conditions including activated ones.

The main result of Chapter II is Theorem II.3.7, which establishes localizability of norms of functionals on dual Sobolev spaces $W^{-1,q}$, $1 \leq q \leq \infty$. This allows one to construct a posteriori error estimators for PDE problems with residuals in $W^{-1,q}$, such that the estimate is reliable (II.1.8) and locally efficient (II.1.10). This result holds under the condition of Galerkin orthogonality (II.3.20). Section II.4.1 investigates situations where the condition is violated. A very simple generalization of the main result for such situations is provided in Theorem II.4.1 while Example II.4.6 links the approach, in a simplified setting, with ℓ^2-estimates of the algebraic residual (which is a very common, but often too crude, practice) and demonstrates that the approach of Theorem II.4.1 might be to crude to be efficient. Theorem II.4.3 together with Example II.4.4 gives a remedy to this problem, which requires a more complicated approach but allows one to recover the local efficiency. Section II.5 gives a numerical example which supports the theoretically obtained results.

In Section III.2 we developed a theory for the PCD preconditioner in infinite-dimensional function spaces. Section III.2.2 investigated conditions under which the PCD operator is guaranteed to be well-defined and invertible on appropriate spaces and provided uniform estimates for its norms and spectrum. A novel aspect of the approach was the relaxation of the requirements on regularity and divergence of the wind. A very important observation about the structure of the preconditioned Schur complement appears in (III.2.54), i.e., that the preconditioned Schur complement $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ is a compact perturbation of the Stokes Schur complement S^{∞}, which is a positive self-adjoint operator. Furthermore, in Section III.2.4 it is shown that the perturbation is of $(6+\epsilon)$-Schatten class and that this implies that the spectrum of $SX_{\alpha,\mathbf{w},\Gamma}^{-1}$ accumulates at the spectrum of S^{∞} with the rate $6+\epsilon$. In Section III.2.5 we discuss the implications of this for the convergence of the GMRES method. Section III.2.6 discusses the relation of the two PCD variants and of the boundary conditions imposed in the definition of the PCD operators. Section III.3.1 provides a methodology for the construction of discrete PCD operators for a broad class of pressure discretizations, including the inflow-outflow situation. The main results are Theorems III.3.2 and III.3.3, which ensure invertibility of and a priori bounds on the discrete PCD operators under appropriate conditions. The subsequent sections then derive particular forms of the PCD operator for specific discretizations. In Section III.3.5 we elaborate on some aspects of previously published accounts and compare these to our results.

Appendix III.B, which is of independent interest in the theory of Krylov subspace methods, provides a new result regarding the convergence of the GMRES method. In particualar, it is shown that compact pertubations of certain operators for which GMRES exhibits contractive convergence are subject to asymptotically contractive convergence with the same contraction factor. Furthermore, if the compact pertubation belongs to some p-Schatten class, the exponent p gives the rate at which this asymptotic behavior is approached.

List of publications

[1] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. Rognes, and G. Wells. “The FEniCS Project Version 1.5”. In: *Archive of Numerical Software* 3.100 (2015). DOI: 10.11588/ans.2015.100.20553.

[2] J. Blechta, J. Málek, and K. Rajagopal. “On the classification of incompressible fluids and a mathematical analysis of the equations that govern their motion”. Submitted. 2019. URL: https://arxiv.org/abs/1902.04853.

[3] J. Blechta, J. Málek, and M. Vohralík. “Localization of the $W^{-1,q}$ norm for local a posteriori efficiency”. In: *IMA J. Numer. Anal.* (Mar. 2019). DOI: 10.1093/imanum/drz002.

[4] M. Řehoř, J. Blechta, and O. Souček. “On some practical issues concerning the implementation of Cahn–Hilliard–Navier–Stokes type models”. In: *International Journal of Advances in Engineering Sciences and Applied Mathematics* (2016), pp. 1–10. DOI: 10.1007/s12572-016-0171-4.